LEVERAGING LARGE SCALE BEEF CATTLE GENOMIC DATA TO IDENTIFY THE ARCHITECTURE OF POLYGENIC SELECTION AND LOCAL ADAPTATION

A Dissertation Presented to the Faculty of the Graduate School
at the University of Missouri-Columbia

In Partial Fulfillment of the Requirements
for the Degree
Doctor of Philosophy

by
TROY ROWAN
Dr. Jared E. Decker,
Dissertation Advisor

DECEMBER 2020

© Copyright by Troy Rowan 2020

All Rights Reserved

APPROVAL PAGE

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

LEVERAGING LARGE SCALE BEEF CATTLE GENOMIC DATA TO IDENTIFY THE ARCHITECTURE OF POLYGENIC SELECTION AND LOCAL ADAPTATION

Presented by Troy Rowan, a candidate for the degree of Doctor of Philosophy, and hereby certify that in their opinion it is worthy of acceptance.

Dr. Jared E. Decker, Animal Science, UMC

Dr. Robert D. Schnabel, Animal Science, UMC

Dr. Jeremy F. Taylor, Animal Science, UMC

Dr. J. Chris Pires, Biological Science, UMC

Dr. Wesley C. Warren, Animal Science, UMC

DEDICATION

I can't say enough thank yous to Harly Durbin, my favorite co-worker, roommate, colleague, and best friend. You've been the person that has helped me with every problem: scientific or not, big or small. I look forward to life full of further collaborations of all kinds. To the pups, Nadia and Cupcake: bork woof woof bork. Thank you to my entire family (Lightfoots, Rowans, Durbins, et al.), especially my little brother Kyle, your support, encouragement, and love has made this possible. Finally, the biggest thank you my heart can muster to the people that have supported me since day one, my parents, Kurt and Theresa Rowan. Thank you for always pushing me, supporting me, and most importantly loving me unconditionally. The sacrifices that you all made to make this possible do not go unnoticed.

This work is dedicated to the four most important men in agriculture that I know: David Hebbert, Harold Rowan, Vern Lightfoot, and Kurt Rowan. While my future won't be tending the same land or cows as you all, I hope to leave as equally large impact on agriculture, the beef industry, and everyone that I interact with as you all did.

ACKNOWLEDGEMENTS

I feel like I could write 200 pages of acknowledgements because my last 23 years of education has been a complete team effort. I was lucky to start my academic career out in what I realize now was an exceptional small-town school. I credit my 5th grade teacher (and favorite neighbor) Betty Brummett with jump-starting my love of science, and my other science teachers for deepening that passion (Rodney Vanderheiden, Jennifer Burn, and Neil Hall) throughout my time at Bedford Community Schools. I realize now that communication skills in science are every bit as important as technical science know-how. I owe so much of my early development in those areas to the Speech and Drama of Bedford Schools (SADOBS) program, particularly to Deb Ritchie, Dee Rankin, and Carl Rankin. During my undergraduate studies at Creighton University, I was lucky to have an incredible set of research mentors (Dr. Karin van Dijk, Dr. Mike McConnell, and Dr. Carol Fassbinder-Orth) that deepened my interest in academic science and research. They prepared me exceptionally well for my Ph.D. program.

I am grateful Dr. Cliff Lamb for pointing me towards the University of Missouri's Animal Genomics program as a potential academic home, it has been the ideal spot for me. Thank you to Dr. Jared Decker, Dr. Bob Schnabel, and Dr. Jerry Taylor for being incredible mentors. The three of you have helped me develop a wide range of technical skills as well preparing me to navigate the academic world. Jared, thank you for taking a chance on a kid with no coding skills or quantitative/population genetics background, but a real passion for cows. You have allowed me to work independently and creatively approach a number of unique and interesting research questions. Bob, you always make me think deeply about each step of the scientific process. This habit that you've helped me develop will serve me well for the rest of my career. I’m also grateful to Dr. Chris Pires and Dr. Wes Warren for serving as members of my committee and important mentors throughout my time at Mizzou. I was also lucky to share an office with some of the brightest and kindest graduate students and postdocs. Dr. Jesse Hoff was critical in helping me get off the ground during my first year at Mizzou and has remained a close friend and colleague. I learned as much from Jesse on our afternoon walk breaks than in most of my classes! Thank you to all of the other members of the MU Genomics team that made my PhD special: Dr. Tamar Crum, Sara Nilson, Dr. Ruben Buckley, Dr. Lynsey Whitacre, Dr. Camila Urbano-Braz, Will Schaffer, Jenna Kalleberg, Esdras Tuyishimire, and Caleb Grohmann.

Thank you to all of the other folks that made Mizzou Animal Science a special place to work. Thank you to Lauren Ciernia for being a constant source of baked goods, unwavering friendship, and being a perfect example of channeling a passion for science into your work every day. Jade Cooper, you are also an incredible example of hard-work, and one of the most supportive people know. Thank you to the exceptional administrators that made my everyday life exponentially easier: Lena, Becki, Debbie, Gloria, and Rob. A special thank you to my friend Elaine, thank you for taking care of us, giving me grief, and lots of laughs over the last 4 ½ years. Thank you to Dr. Bill Lamberson for your leadership in the department and for the opportunity to teach Animal Breeding, that was a wonderful experience! I was lucky to have so many friends, collaborators, and colleagues from the “other side of campus” that broadened my exposure to genetics outside of cows. Thank you especially to Dr. Sarah Turner-Hissong

(Alex and Sadie) and Dr. Makenzie Mabry (Dr. Paul Blischak), and Paul Petrowski for your friendship, support, and collaboration.

I was lucky to spend six months of 2019 in Dr. John Hickey's group at the Roslin Institute in Edinburgh, Scotland. I appreciate John's willingness to host me, as well as the mentorship and friendship of the whole AlphaGenes team, especially Dr. Gregor Gorjanc, Romana Gorjanc, Dr. Ivan Porcnic, Dr. Chris Gaynor, Dr. Martin Johnsson, Dr. Owen Powell, and of course The Brazilians! This was an amazing visit, scientifically, culturally, and socially. Thank you all for making my time across the pond so special.

My Ph.D. project would not have been possible without collaborations with three breed associations; the American Simmental Association, the American Gelbvieh Association, and the Red Angus Association of America. Thank you to the breed associations and their producers. The ability to utilize your genomic data has saved me years of DNA collection! In particular thank you to Dr. Jackie Atkins, Jordan Bowman, and Ryan Boldt for all of your assistance over the past 4 ½ years.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS …………………………………………………………… ii

LIST OF FIGURES…………………………………………………………….. x

LIST OF TABLES …………………………………………………………… xiii

ABSTRACT ……………………………………………………………….... xvi

CHAPTER 1………………………………………………………………….. 1

LEVERAGING COMMERCIALLY-GENERATED GENOMIC DATA FOR BASIC SCIENCE IN BEEF CATTLE POPULATIONS…………………………… 1

- **Genotype imputation…………………………………………………..... 1**
- **Polygenic selection………………………………………………………… 4**
- **Local adaptation……………………………………………………….... 6**

CHAPTER 2………………………………………………………………….. 9

A MULTI-BREED REFERENCE PANEL AND ADDITIONAL RARE VARIATION MAXIMIZE IMPUTATION ACCURACY IN CATTLE…………… 9

- **Abstract…………………………………………………………………..... 10**
- **Background………………………………………………………………. 11**
- **Methods………………………………………………………………….... 14**
  - Quality control and filtering …………………………………………… 15
  - Creating the imputation test set………………………………………… 15
  - Building phasing and imputation reference panels…………………….. 16
  - Measures of imputation accuracy……………………………………… 18
- **Results…………………………………………………………………….. 20**
  - Imputation accuracy metrics…………………………………………… 20

Impact of the breed representation in the reference panel on imputation accuracy …………………….……………………………………….. 24
Error profiles and regions of low imputation accuracy ………………... 27
Discussion ……………………………………………………………... 28
Imputation accuracy metrics…………………………………………… 28
Impact of the F250 assay on imputation of rare variants……………….. 29
Multi-breed vs. within-breed imputation reference panels …………….. 30
Breed representation in the reference panel ……………………………. 32
Starting assay marker number ………………………………………….. 33
Conclusions ……………………………………………………………... 33
Acknowledgements …….……………………………………………….... 34
Author Contributions …………………………………………………… 35
CHAPTER 3………………………………………………………………………… 58
POWERFUL DETECTION OF POLYGENIC SELECTION AND ENVIRONMENTAL ADAPTATION IN U.S. BEEF CATTLE ………………….. 58
Abstract……………………………………………………………….... 59
Author Summary ……………………………………………………….. 60
Introduction …………………………………………………………… 60
Results………………………………………………………………….. 62
Detecting ongoing polygenic selection with Generation Proxy Selection Mapping (GPSM)……………………………………………………. 62
Detecting environmental adaptation using envGWAS ………………… 68
Discrete ecoregion envGWAS ……………………..……………….... 70

Continuous environmental variable envGWAS ……………………….. 72

Identifying loci undergoing region-specific selection with GPSM ecoregion meta-analysis ……………………………………………….. 74

Discussion ………………………………………………………………... 76

Materials and Methods …….……………………………………………... 79

Genotype data ………….…………………………………………… 79

Generation Proxy Selection Mapping (GPSM) ……………………... 80

Birth date variance component analysis ……………………………….. 81

Environmental data ……………………………………………………….. 81

Environmental Genome-wide Association Studies (envGWAS) ……… 82

GPSM meta-analyses ………………………………………………….. 83

Gene and tissue set enrichment analyses ……………………………… 84

Acknowledgements …….……………………………………………... 85

Supplemental Text ………………………………………………………... 135

Motivation for mapping polygenic selection ………………………… 135

Motivation for mapping local adaptation …………………………….. 136

Simulations ………………………………………………………… 138

GPSM detects signatures of polygenic selection across and within three populations of U.S. Beef Cattle ………………………………………. 142

Principal Component Analysis ……………………………………….. 145

envGWAS identifies adaptive pathways and processes ……………… 146

Tissue set enrichment analysis ……………………………………….. 149

GPSM meta-analysis identifies loci responding to selection in an ecoregion-specific manner …………………………………………… 152

CHAPTER 4………………………………………………………………...154

POWERFUL DETECTION OF POLYGENIC SELECTION AND ENVIRONMENTAL ADAPTATION IN U.S. BEEF CATTLE …………………..154

Abstract………………………………………………………………...155

Introduction …………………………………………………………….. 156

Results………………………………………………………………… 158

Quantitative genetic signals of polygenic selection…………………… 158

Generation Proxy Selection Mapping (Red Angus) ………………….. 159

Generation Proxy Selection Mapping (Simmental) …………………... 162

Ongoing selection at functional loci ………………………………….. 164

GPSM identifies breed-specific balancing selection in KHDRBS2 regulatory regions …………………………………………………….. 166

Sweep mapping identifies known and novel selected loci …………… 167

Ongoing polygenic selection acts on networks and pathways to drive complex trait changes ………………………………………………… 171

Discussion ……………………………………………………………. 172

Conclusions …………………………………………………………... 175

Methods ……….……………………………………….......................... 175

Genotype data and imputation ……………………………………....176

Generation proxy phenotypes ……………………........................... 176

Generation Proxy Selection Mapping (GPSM) ………………………. 177

Haplotype-based scans for selection ………………………………….. 178

Site frequency spectrum (SFS) scan for selection ……………………. 179

Gene &QTL annotation and enrichment ……………………………… 179

REFERENCES …………………………………………………………………. 203

VITA ………………………………………………………………………………. 229

LIST OF FIGURES

Figure **Page**

2.1 Minor allele frequency spectra for three commercially available assays………. 37

2.2 Imputation quality score compared to other measures of accuracy ……………. 38

2.3 Imputation accuracy across minor allele frequencies using various references .. 39

2.4 Individual imputation accuracy changes when using breed vs. composite imputation reference panels ………………………………………………… 40

2.5 Per-individual accuracy in different reported breeds ………………………….. 41

2.6 Scatter plot relating within-reference relatedness and imputation accuracy ..…. 42

2.7 Effects of assay density on imputation accuracy across MAF spectrum ………. 43

2.8 Impact of starting assay marker assay number on per-individual imputation accuracy………………………………………………………………………… 44

2.9 Regions of low imputation accuracy across the genome ………………………. 45

2.10 Representation of imputation accuracy testing scheme ……………………….. 46

2.11 Schematic representation of "one-round" vs. "two-round" imputation ……….. 47

2.12 Imputation quality across MAF spectrum ……………………………………… 48

2.13 Impact of genetic similarity with reference on imputation accuracy in three breeds ……………………………………………………………………………….. 49

3.1 Simulated allele frequency trajectories and a model overview ………………... 86

3.2 Generation Proxy Selection Mapping (GPSM) panel ………………………… 88

3.3 Environmental GWAS (envGWAS) panel …………………………………….. 90

3.4 Meta-analysis of within-ecoregion selection in Red Angus …………………… 92

3.5 Distribution of continuous birth date in populations…………………………… 93

3.6 Manhattan plots for discrete envGWAS in Red Angus ………………………... 94

3.7 Manhattan plots for discrete envGWAS in Simmental ………………………… 96

3.8 Manhattan plots for discrete envGWAS in Gelbvieh ………………………….. 98

3.9 PCA (PCs 1-8) for Red Angus, Simmental, and Gelbvieh populations ……….. 99

3.10 Manhattan plots for continuous envGWAS in Red Angus …………………… 100

3.11 Manhattan plots for continuous envGWAS in Simmental …………………… 101

3.12 Manhattan plots for continuous envGWAS in Gelbvieh ……………………... 102

3.13 PM-plots and region-specific allele frequency changes for selected Red Angus GPSM meta-analysis SNPs …………………………………………………… 103

3.14 PM-plots and region-specific allele frequency changes for selected Simmental GPSM meta-analysis SNPs …………………………………………………… 105

3.15 PM-plots and region-specific allele frequency changes for selected Gelbvieh GPSM meta-analysis SNPs …………………………………………………… 107

4.1 Red Angus GPSM Manhattan Plots ………………………………………... 181

4.2 Comparison of Red Angus GPSM signatures with sweep mapping methods …182

4.3 Simmental GPSM Manhattan Plots ………………………………………... 184

4.4 Chromosome 23 GPSM and functional annotation …………………………... 185

4.5 Birth date distributions ……………………………………………………... 187

4.6 Raw birth date GREML residual and breeding value Q-Q plots ………………188

4.7 Transformed and truncated birth date GREML residual and breeding value Q-Q plots …………………………………………………………………………… 189

4.8 811K SNP GPSM Manhattan plots comparing effects of transformations and subsetting ……………………………………………………………………... 190

4.9 Changing breed composition of Simmental ancestry over time ……………… 191

4.10 Red Angus selective sweep mapping Manhattan Plots ……………………….. 192

4.11 Simmental selective sweep mapping Manhattan Plots ……………………….. 193

LIST OF TABLES

Table **Page**

2.1 Variant counts for masked genotypes in testing individuals ...………………… 50

2.2 Breed representation in the composite reference after removing test set ……… 51

2.3 Per-variant mean imputation accuracy by MAF bins ………………………….. 53

2.4 Mean imputation quality by MAF bin for HD-specific markers imperfectly imputed using the “two-round” imputation method …………………………… 54

2.5 Individual imputation accuracies by breed …………………………………..... 55

2.6 Per-variant imputation accuracy based on starting assay density ……………… 56

2.7 Per-individual imputation accuracy based on starting assay density …………... 57

3.1 Variation in birth date explained by three classes of SNPs …………………... 109

3.2 GPSM and envGWAS gene dropping simulation results …………………….. 110

3.3 GPSM stochastic simulation results …………………………………………... 111

3.4 Dataset summary statistics ……………………………………………………. 112

3.5 Proportion of variation in birth date explained by markers in GPSM ………... 113

3.6 Summary statistics for significant GPSM SNPs ……………………………… 114

3.7 Allele frequency change per generation summary statistics ………………….. 115

3.8 Significant GPSM variants identified in at least two populations ……………. 116

3.9 Gene enrichment analysis of GPSM candidate genes ……………………….. 118

3.10 TissueEnrich results for GPSM gene sets …………………………………….. 119

3.11 Tissue enrichment analysis results from pSI using GPSM candidate genes …. 120

3.12 Counts of animal per ecoregion by population ……………………….......... 121

3.13 Univariate REML PVE estimates for continuous environmental variables …...122

3.14 Univariate REML PVE estimates for ecoregion membership ………………... 123

3.15 Candidate genes for discrete ecoregion associations in Red Angus ………….. 124

3.16 envGWAS significant SNPs and candidate genes ……………………………. 126

3.17 envGWAS candidate genes …………………………………………………… 127

3.18 Gene enrichment analysis of envGWAS candidate genes ……………………. 130

3.19 TissueEnrich results for envGWAS candidate genes ………………………… 131

3.20 Tissue enrichment results from pSI using envGWAS candidate genes ……… 132

3.21 Brain region and cell-type enrichments of envGWAS candidate genes ……… 133

3.22 *C. elegans* cell-type enrichments for envGWAS candidate genes …………… 134

4.1 Genetic architecture of GPSM COJO associations …………………………… 194

4.2 Genetic architecture of nSL and RaiSD windows ……………………………. 195

4.3 GREML estimates of proportion variance explained for various statistical transformations to birth date in Red Angus ………………………………….. 196

4.4 GREML estimates of proportion variance explained for various statistical transformations to birth date in Simmental ………………………………….. 197

4.5 Number of significant SNPs identified by different transformations or subsets of birth date at different significance thresholds in Red Angus ....…………… 198

4.6 Number of significant SNPs identified by different data subsets at different significance thresholds in Simmental ....…………………………………... 199

4.7 Significant COJO SNPs from full Red Angus GPSM analysis ………………. 200

4.8 Significant COJO SNPs from young Red Angus GPSM analysis ……………. 200

4.9 Significantly enriched QTL types near full Red Angus dataset GPSM COJO SNPs …………………………………………………………………………………… 200
4.10 Significantly enriched QTL types near young Red Angus dataset GPSM COJO SNPs……………………………………………………………………………… 200
4.11 Significant COJO SNPs from full Simmental GPSM analysis ………………… 200
4.12 Significant COJO SNPs from purebred Simmental GPSM analysis …………. 201
4.13 Significantly enriched QTL types near full Simmental dataset GPSM COJO SNPs …………………………………………………………………………………… 201
4.14 Significantly enriched QTL types near purebred Simmental dataset GPSM COJO SNPs……………………………………………………………………………… 201
4.15 Red Angus outlier nSL windows and associated genes ………………………. 201
4.16 Red Angus outlier RAiSD windows and associated genes …………………… 201
4.17 Simmental outlier nSL windows and associated genes ………………………. 202
4.18 Simmental outlier RAiSD windows and associated genes …………………… 202

LEVERAGING LARGE SCALE BEEF CATTLE GENOMIC DATA TO IDENTIFY THE ARCHITECTURE OF POLYGENIC SELECTION AND LOCAL ADAPTATION

Troy Rowan

Dr. Jared Decker, Dissertation Supervisor

ABSTRACT

Since the invention of the first array-based genotyping assay for cattle in 2008, millions of animals have been genotyped worldwide. Leveraging these genotypes offers exciting opportunities to explore both basic and applied research questions. Commercial genotyping assays are of adequate variant density to perform well in prediction contexts but are not sufficient for mapping studies. Using reference panels made up of individuals genotyped at higher densities, we can statistically infer the missing variation of low-density assays through the process of imputation. Here, we explore the best practices for performing routine imputation in large commercially generated genomic datasets of U.S. beef cattle. We find that using a large multi-breed imputation reference maximizes accuracy, particularly for rare variants. Using three of these large, imputed datasets, we explore two major population genetics questions. First, we map polygenic selection in the bovine genome, using Generation Proxy Selection Mapping (GPSM). This identifies hundreds of regions of the genome actively under selection in cattle populations. Using a similar approach, we identify dozens of genomic variants associated with environments across the U.S., likely involved local adaptation. Understanding the genomic basis of local adaptation in cattle will enable select and breed cattle better suited to a changing climate.

CHAPTER 1

LEVERAGING COMMERCIALLY GENERATED GENOMIC DATA FOR BASIC SCIENCE IN BEEF CATTLE POPULATIONS

In a landmark paper in 2001, Meuwissen, Hayes, and Goddard laid the foundation for the most important development in plant and animal breeding this millennium [1]. Their idea to use high-density SNP genotypes across the genome to more accurately estimate breeding values would change how breeding programs and whole industries operated. It would be another seven years until their theory could be realized in practice, as no high-density genomic assays for agricultural species existed at the time. The development of the BovineSNP50 in 2008 ushered in a new era of animal breeding and allowed the theory of Meuwissen, Hayes, and Goddard to be realized and implemented in genetic evaluations [2,3]. The impact of genomic selection was immediate. The ability to obtain accurate predictions of genetic merit on young animals did away with dairy progeny testing programs virtually overnight [4]. This shortened generation intervals and led to a massive acceleration in genetic gain [5]. In dairy cattle, the adoption of genomic technologies was instantaneous, and now over 700,000 animals are genotyped per year in the United States [6]. Genomic selection uptake was slower in the beef industry, but it has increased substantially in recent years, leading to millions of genotyped animals across breeds.

Genotype Imputation

In most cases, commercially-generated assays genotype between 7,000 and 70,000 SNPs [7]. These assays are designed to genotype SNPs with moderate-to-high minor allele frequency (MAF) that are evenly spaced throughout the genome. The assays are of sufficient density to effectively define relationships between individuals for use in routine genomic prediction [8]. Adding more common markers tends to provide only subtle increases in prediction accuracy [9]. However, these assays are not of sufficient density to precisely map quantitative trait loci in genome-wide association studies (GWAS). As we attempt to leverage commercially-generated data for use in mapping studies, increasing the marker density is a crucial first step in adding utility to the data. In populations with large reference sets of individuals with high-density or full sequence genotypes, missing genotypes that occur when using low-density assays can be statistically inferred through imputation [10]. In the context of reusing commercially-generated data or designing genome-wide association studies (GWAS), imputation allows the genotyping of large numbers of individuals (high power) without sacrificing marker resolution (high numbers of SNPs).

It is important to note that the utility of imputation in mapping studies or genomic prediction depends on its accuracy. Imputation accuracy is a function of multiple factors, some of which are algorithmic, and others that are dependent on the genetic architecture of the dataset and reference haplotypes. At its core, imputation is a pattern matching process, and accurate imputation of the missing alleles within a haplotype requires that the haplotype resides within the set of high-density reference haplotypes. In most of the major cattle breeds, large numbers of individuals have been genotyped using high-density assays and/or have sequenced at high coverage. Consequently, a large pool of haplotypes exists for use as reference sets in imputation to either high-density chip-level (~800,000 SNPs) or full sequence (tens of millions of SNPs). In cattle, imputation directly from low-density to full sequence genotypes results in substantial drops in accuracy, meaning that commercial assays first need to be imputed to a high-density chip level [11] prior to imputation to the level of whole-genome resequencing data.

In Chapter 2 [7], we examine the effects of multiple non-algorithmic alterations on imputation accuracy to the high-density chip level. We test the impact of the imputation reference panel on downstream imputation accuracy using down-sampled genotypes from animals' true high-density genotypes. The down-sampled data allowed us to categorize and quantify the magnitude of imputation errors at the variant and individual level. By using a large multi-breed reference panel of high-density reference haplotypes, we demonstrate significant gains in imputation accuracy, particularly for rare variants. The multi-breed reference panel also increases the imputation accuracies of genotypes obtained for crossbred or admixed animals. This is especially important as much of the commercially-generated data that we receive from cattle breed associations includes crossbred animals, making it difficult to pinpoint the exact breed ancestry of individuals for which to assemble a within-breed imputation reference panel. The optimal imputation strategy and reference panel identified in Chapter 2 are implemented as a Snakemake [12] pipeline that is routinely used to perform genotype imputation for the Mizzou Animal Genomics group. The genomic data used to map polygenic selection and local adaptation in Chapters 3 and 4 were imputed based on the workflow described in Chapter 2. Ongoing improvements to this pipeline, both in software and reference panel content, will continue to increase the accuracy of imputation, and thus the usefulness of commercially-generated beef cattle genotypes for answering basic and applied research questions.

Leveraging high-density, imputed genotypes from tens of thousands of beef cattle genotyped over time and across diverse landscapes, we explore two major questions in population genetics in Chapter 3. First, we use genome-wide association study (GWAS) models to detect variants undergoing polygenic selection in a method called Generation Proxy Selection Mapping (GPSM) [13,14]. Next, we use these same GWAS models to identify genomic associations with an individual's environment as means for mapping variants associated with local adaptation.

Polygenic Selection

In both wild and domestic populations, selection occurs on phenotypes that increase fitness, or economically important traits, respectively. This selection on certain phenotypes changes the frequencies of the genetic variants that underlie variation in these traits [15]. When selected traits are simple or Mendelian in nature, controlled by relatively few loci, allele frequencies change rapidly. For these traits, beneficial causal loci are rapidly "swept" to fixation in the population, causing frequency changes at neighboring neutral sites due to linkage disequilibrium [16]. Much work has been done to map these selective sweeps in humans, cattle, and numerous other species [17–19]. It is becoming increasingly apparent that in most populations, most selection is polygenic, acting on complex traits, controlled by hundreds or thousands of loci spread throughout the genome [20]. The result is that polygenic selection can cause large shifts in a population's mean phenotype without creating detectable shifts in allele frequency at any single locus [21,22].

Most traits under selection in cattle are complex in nature. While sweep mapping has identified numerous locations throughout the genome that have undergone strong selection in the distant or intermediate past [23,24], it remains largely unknown how polygenic selection impacts the genome over short timescales. Commercially-sourced cattle genotype data presents an opportunity to understand the genomic changes due to ongoing polygenic selection, because: 1) it is distributed over time (widespread genotyping since 2008 and genotypes on early founder individuals), 2) the selection goals in populations are well-understood, 3) there is an enormous statistical power (thousands of genotyped individuals) to detect small allele frequency shifts. Decker et al. (2012) [13] developed a method to map polygenic selection in populations with temporally-stratified genomic data. Their method, Generation Proxy Selection Mapping (GPSM), utilizes genome-wide association study (GWAS) models to detect allelic associations with an individual's generation, or some proxy for generation. In addition to being scalable to large numbers of individuals and marker tests, the use of GWAS models with a genomic relationship matrix (GRM) controls for population and family structure that exist in the dataset. These significant associations between an allele and an individual's generation insinuate that a locus is undergoing changes in frequency greater than expected due to drift [25].

In Chapter 3, we extend the GPSM method to three large genotyped beef cattle populations [14]. Additionally, we perform simulations to validate the method and infer how its power to detect selection varies under certain genomic architectures, effective population sizes, and strengths of selection. In addition to increasing the number of genotyped samples compared to Decker et al. (2012), we use significantly higher-density imputed genotypes (from methods described in Chapter 2). This allows the more precise mapping of selected variants and identification of their positional candidate genes.

Chapter 4 further expands on this approach, using two cattle datasets of 50,000 and 75,000 individuals imputed to 11 million variants. This is likely one of the largest non-human selection mapping datasets to date. Using sequence-imputed genotypes allows for the functional annotation of GPSM SNPs and a deeper understanding of the genetic architecture that governs traits being selected in contemporary cattle populations. We demonstrate that GPSM can identify minute allele frequency shifts caused by selection over very short time periods (< 2 generations). We compare GPSM signatures to traditional methods of selection mapping and find minimal overlap. Beyond identifying different selected loci, we show that polygenic selection identified by GPSM does not alter neighboring neutral diversity like the selective sweeps identified by other methods. The GPSM method offers an intriguing complement to other selection mapping approaches when temporally-stratified genotype data are available. It also offers a glimpse at the impacts of producer selection decisions in the beef industry.

Local Adaptation

Beef cattle are the last major livestock species in the United States to reside largely in uncontrolled environments “outdoors”. As pork, poultry, and many dairy operations have moved into highly-controlled indoor environments, beef cattle remain exposed to the full suite of environmental stressors. These can be purely climatic like heat [26] or cold stress, or more complex environmental stressors like water availability [27] or presence of toxic fescue [28]. Gene-by-environment interactions are well-documented in beef cattle populations [29–31], especially for growth traits. Still, the genetic variants underlying GxE interactions are not well understood. Most GxE mapping studies in beef cattle have focused on identifying SNPs with significant growth trait-by-environment interactions [32,33], treating GxE as a quantitative genetic problem. It is also clear that despite the use of national genetic evaluations used to select sires, cattle populations are directly selecting for animals that produce well in their herd's environment. An experiment involving the reciprocal transplant of Line 1 Hereford cattle between Montana and Florida in the 1970s showed that highly-similar genetics selected in different environments experienced massive performance losses when exposed to a novel stressful environment [34]. Though the selection goals in two different environments may be the same, animals selected in each environment utilize different sets of variation to optimize performance in each environmental context. Over time we would expect that the genetic variants that confer local adaptation to animals would exhibit allele frequency differences along environmental gradients (i.e. temperature, precipitation, etc.) and between diverse regions [35,36].

Using a similar approach to GPSM in Chapter 3, we leverage large beef cattle populations distributed across a range of environments in the Continental United States to map DNA variants associated with continuous environmental variables at an animal's birth location (30-year normal temperature, precipitation, elevation), or an individual's membership in one of nine statistically-derived ecoregions [14]. We identified dozens of variant-environment associations, indicating that adaptive alleles do, in fact, exist and their frequencies differ between environments. While the associations were largely different between datasets from three different breeds, the biological pathways in which their positional candidate genes operate, were similar. The enriched pathways shared between breeds were overwhelmingly involved in neural development and signaling, despite being driven by entirely different loci and candidate genes. Further, using GPSM within environmental regions, we identified ongoing ecoregion-specific selection. While we might expect that producers recurrently select better-adapted cattle, the reality is that the use of national genetic evaluations and artificial insemination in beef cattle is eroding the existing allele frequency differences that exist between populations.

CHAPTER 2

A MULTI-BREED REFERENCE PANEL AND ADDITIONAL RARE VARIATION MAXIMIZE IMPUTATION ACCURACY IN CATTLE

Troy N. Rowan [1], Jesse L. Hoff [1], Tamar E. Crum [1], Jeremy F. Taylor[1], Robert D. Schnabel [1,2], and Jared E. Decker [1,2*]

1 Division of Animal Sciences, University of Missouri, Columbia, Missouri 65211, USA

2 Informatics Institute, University of Missouri, Columbia, Missouri 65211, USA

*Corresponding author

Abstract

Background: During the last decade, the use of common-variant array-based single nucleotide polymorphism (SNP) genotyping in the beef and dairy industries has produced an astounding amount of medium-to-low density genomic data. Although low-density assays work well in the context of genomic prediction, they are less useful for detecting and mapping causal variants and the effects of rare variants are not captured. The objective of this project was to maximize the accuracies of genotype imputation from medium- and low-density assays to the marker set obtained by combining two high-density research assays (~ 850,000 SNPs), the Illumina BovineHD and the GGP-F250 assays, which contains a large proportion of rare and potentially functional variants and for which the assay design is described here. This 850 K SNP set is useful for both imputation to sequence-level genotypes and direct downstream analysis.

Results: We found that a large multi-breed composite imputation reference panel that includes 36,131 samples with either BovineHD and/or GGP-F250 genotypes significantly increased imputation accuracy compared with a within-breed reference panel, particularly at variants with low minor allele frequencies. Individual animal imputation accuracies were maximized when more genetically similar animals were represented in the composite reference panel, particularly with complete 850 K genotypes. The addition of rare variants from the GGP-F250 assay to our composite reference panel significantly increased the imputation accuracy of rare variants that are exclusively present on the BovineHD assay. In addition, we show that an assay marker density of 50 K SNPs balances cost and accuracy for imputation to 850 K.

Conclusions: Using high-density genotypes on all available individuals in a multi-breed reference panel maximized imputation accuracy for tested cattle populations. Admixed animals or those from breeds with a limited representation in the composite reference panel were still imputed at high accuracy, which is expected to further increase as the reference panel expands. We anticipate that the addition of rare variants from the GGP-F250 assay will increase the accuracy of imputation to sequence level.

Background

High-density single nucleotide polymorphism (SNP) genotyping has driven rapid improvements in rates of genetic progress in livestock populations [5,37,38]. To increase the predictive capabilities of genomic prediction models further, the discovery of functional variants has become increasingly important. Although many large-effect or Mendelian variants that control important phenotypes in cattle have been identified [39–43], the identification of moderate and small effect quantitative trait nucleotides (QTN) and other causal variants has proven challenging [44,45]. Early genome-wide association studies (GWAS) that focused on the detection of these variants were often forced to choose between the density of the SNP array (number of SNPs genotyped) and statistical power (number of individuals genotyped). Imputation, the use of statistical models, and a reference set of haplotypes to infer missing genotypes, allows researchers to genotype large numbers of individuals at relatively low-density and impute their genotypes to high-density or even millions of SNPs from whole-genome resequencing data [46–48].

Low- to medium-density common variant SNP assays are widely used for genetic evaluation in both beef and dairy cattle. Since the development of the BovineSNP50 (SNP50) BeadChip (Illumina, San Diego, CA) [2] in 2008 and the BovineHD (Illumina, San Diego, CA) array in 2009, more than 3 million dairy cattle in the United States alone have been genotyped using SNP assays that are derived from these progenitor assays [49]. Decker [50] noted the value of these commercially-generated datasets for uses beyond genetic prediction. Although lower-density assays work well for genomic prediction [1,4,49], the effects due to rare variants are not captured and they have a low resolution for the detection of quantitative trait loci (QTL) or causal variants. High-quality imputation allows these datasets to be used to their full potential [11,51,52]. Seabury et al. [53] found that similar trait heritabilities were obtained with 50 K common variant genotypes and 778 K common variant imputed genotypes, but that the former were less powerful for QTL detection. Imputed 778 K genotypes identified 14 putative large effect QTL that were not identified using 50 K genotypes. Using these large publicly-funded or commercially-generated datasets imputed to high-resolution marker densities will increase prediction accuracies, aid in the detection of causal variants, and ultimately increase selection response in cattle [11,54–56].

To use these large datasets to their full potential, the accuracy of imputation must be maximized. The most accurate imputation software packages for cattle [51,57] were typically developed for human studies that were aimed at imputing from a high-density genotype panel to full-genome sequence. As a result, using these programs to impute genotypes directly from low-density to full-genome sequence, even in cattle breeds with high levels of linkage disequilibrium (LD), has been less accurate [11]. A “two-step” imputation strategy, first from a low-density assay (8 K to 70 K variants) to a high-density assay (> 700 K variants) and then from imputed high-density to the sequence level was more accurate than genotypes imputed in “one-step” from low-density to full-genome sequence in both cattle and humans [58,59]. In this study, we consider the first part of the “two-step” imputation processes, because the produced genotypes can be used as input for imputation to full-genome sequence or as an endpoint for a variety of downstream analyses. Regardless of its use, maximizing the accuracy of imputation to high-density genotypes is essential to the success of both approaches.

Initially, SNP assays for cattle were designed with common, evenly spaced markers that would presumably be in LD with causal variants [2]. Whereas these assays have performed well in genomic prediction applications, there is growing interest in including rare variants into predictions [11,46,56,60]. Imputation accuracy has been shown to decline rapidly as minor allele frequencies (MAF) of SNPs decrease, thus increasing the confidence in the imputation of genotypes for rare variants has become a priority. In addition, most studies on optimizing imputation have focused on the imputation of genotypes for purebred animals using closely-related individuals from the same breed. As large numbers of genotypes for unpedigreed crossbred animals have become available, it is necessary to re-evaluate strategies for genotype imputation in these datasets.

This study focuses on maximizing imputation accuracy from several commercially available low-density common variant SNP genotyping assays to a set of high-density variants (850 K), many of which are rare and potentially functional. We test the effectiveness of a large, multi-breed composite reference panel for imputation in several beef and dairy cattle populations that are genotyped with several commercially available common variant SNP genotyping assays. We use both well-established and novel measures of imputation accuracy to categorize precisely the causes of imputation errors. These metrics provide insights for interpretation of imputation performance and define situations in which researchers should be cautious when using imputed variants. In addition, we explore how the starting chip density impacts the accuracy of imputation to 850 K variants. Finally, we introduce and describe the design of the GGP-F250 functional genotyping assay. The GGP-F250 is a tool not only for genotyping numerous functional variants but also for increasing the imputation accuracy of rare variants.

Methods

To identify the best practices for achieving imputation accuracies that approach the error rates of modern SNP genotyping arrays, we compared the impact of altering reference panels and marker numbers in the starting assay when imputing genotypes to the level of the combined Illumina BovineHD (Illumina, San Diego, CA) and GeneSeek Genomic Profiler F250 (GeneSeek. Lincoln, NE) referred to herein as the HD and F250 assays, respectively. The HD assay contains 777,962 evenly spaced variants that have relatively high MAF across many breeds of cattle common to North America. The F250 assay contains 227,234 markers, of which 31,392 are present on the HD assay and included in the assay design for use in imputation, and another 195,842 potentially functional markers, many of which are rare (MAF < 0.1). Due to these rare alleles, the MAF distribution for the F250 assay is more similar to the site frequency spectrum of the bovine genome (Figure 1). Details on the design of the F250 assay are in Additional file 1: Tables S1–S4. In this study, we used 2718 animals that were genotyped with both the

F250 and HD assays, and 25,772 animals that were genotyped with only the F250, and 7218 animals genotyped with only the HD assay.

Quality control and filtering

Prior to sub-setting and masking genotypes for testing, we used the PLINK1.9 software [61] to filter variants and individuals. The SNP positions were based on the ARS-UCD1.2 bovine reference genome assembly [62]. Non-autosomal variants were removed from the data. Variants and individuals with call rates lower than 0.90 were removed from the testing and reference datasets. Because many of the F250 variants are rare, no MAF filter was applied to any of the SNP arrays. Due to the diverse breed composition of the dataset, no Hardy–Weinberg equilibrium filter was applied. PLINK was used to estimate MAF in the filtered dataset for use in all downstream analyses. Two animals were removed due to low genotype call rates. The numbers of remaining variants after filtering for each of the assays in the masked testing set are in Table 1.

Creating the imputation test set

To test the accuracy of imputation, PLINK1.9 [61] was used to down-sample genotypes for 307 animals with both HD and F250 genotypes to the densities found on several commonly used commercial genotyping arrays: SNP50 and GGP-LD, GGP-90KT, GGP-HDv3, and GGP-ULD (all from GeneSeek, Lincoln, NE), which were then imputed to the combined high-density research chips (~ 850 K SNPs). The process of sampling and masking testing genotypes is described visually in Additional file 2: Figure S1. All tested commercial assays possess SNPs that are largely derived from the SNP set of the HD assay (see Additional file 3: Table S5).

A maximum of 50 individuals per breed that were genotyped with both the HD and F250 assays, were randomly chosen and masked to represent various commercial chip densities for testing imputation accuracy (Table 2). All test set individuals had their breed-composition estimated by the CRUMBLER pipeline [63]. To avoid depleting the reference panel of breeds with small numbers of research assay genotypes, no more than 50% of a breed's F250 or HD genotyped animals were removed for testing. The remainder of the HD and F250 genotypes were used in the composite reference panel (Table 2). Due to the unequal representation of breeds in the test dataset, we created three separate datasets for testing different aspects of our imputation pipeline. The first dataset, ALL, used all 307 masked individuals that passed genotype call rate filtering. Because some of the indicine breeds used in our testing dataset were not adequately represented in the imputation reference panel, or their testing dataset sample sizes were not sufficiently large to draw meaningful conclusions, we created a test dataset, TAUR, which comprised only Bos taurus animals, i.e. 281 Angus, Gelbvieh, Hereford, Holstein, Limousin and Simmental individuals. Finally, we used a test dataset, GEL that included 49 Gelbvieh individuals, to compare the accuracy of a within-breed imputation reference to the composite reference.

Building phasing and imputation reference panels

After removing 307 individuals for testing, the remaining 28,183 F250 and 9629 HD genotyped reference individuals (Table 2) were merged in PLINK and then phased with Eagle 2.4 [64]. Missing genotypes inferred by Eagle were removed with the bcftools program [65] such that only the phased, directly genotyped markers remained.

The within-breed imputation reference panel consisted of 265 and 514 Gelbvieh individuals that were genotyped with the HD and F250 assays, respectively. These reference individuals had their genotypes merged and phased, and the inferred genotypes were removed separately for each assay. Reciprocal F250/HD imputation analyses performed with Minimac3 were used to fill in missing genotypes in the reference panel.

Phasing and imputation

Reference-based phasing was performed for 307 individuals with masked genotypes in Eagle using 9629 individuals with pre-phased HD assay genotypes as the reference haplotypes. To perform "one-round" imputation, phased assays were imputed against the complete imputed 850 K SNP composite reference panel using Minimac3 [66]. The reference panel for the "one-round" imputation process was created by imputing missing HD markers for individuals genotyped on the F250 assay, and missing genotypes for F250 markers for individuals genotyped on the HD assay with Minimac3 (see Additional file 4: Figure S2). Here, the reference panel contained both observed and imputed genotypes.

For "two-round" imputation, two separate imputation steps were performed to reach the 850 K SNP density (see Additional file 4: Figure S2). In each step, only observed genotypes served as HD and F250 references, respectively (no imputed genotypes in reference). First, the testing individuals with masked and phased genotypes were imputed to HD density (759,329 SNPs), and then a second imputation step was performed that inferred genotypes for markers present on the F250, but not on the HD

(122,181 SNPs) assay. Both imputation methods resulted in a total number of 835,947 variants, of which 835,926 segregated in the "one-round" CR panel and 835,933 in the "two-round" CR panel.

For the within-breed imputation, 49 Gelbvieh animals, all of which were present in the multi-breed testing set, which had been genotyped with both the F250 and HD assays, were masked to SNP50 density. Genotypes for these individuals were phased using Eagle along with 1113 additional Gelbvieh individuals genotyped with the SNP50 assay. This is representative of phasing strategies that involve a large number of individuals that have been genotyped using lower density assays. Phased genotypes were imputed against the breed-specific Gelbvieh reference (BR) panel.

Measures of imputation accuracy

Imputation accuracy was measured for both individuals and variants within each imputation scenario. By coding alternate allele counts as 0, 1, and 2 (for AA, AB, and BB genotypes, respectively), both Pearson's correlation coefficient (r) and count-based metrics could be used to evaluate the imputation accuracy for each variant and individual. Pearson's correlation coefficients for individuals were calculated in two ways. First using unscaled, raw genotype values and then using genotype values centered by the variant's MAF in the entire set of research assays (all HD, F250, reference and testing). To center genotypes, twice the variant's MAF was subtracted from the raw genotype value. For both methods, r values were calculated and compared.

Although simple concordance (i.e., "correct/incorrect") measures of accuracy are valuable, they overestimate the quality of imputation at low MAF and are ambiguous as to the nature of the error that created an incorrectly imputed genotype. Rather than concordance rate, an imputation quality score (IQS) [67] was calculated for each variant. The IQS calculates concordances that are adjusted for the chance that an imputed genotype could be correctly guessed. This statistic provides similar conclusions to correlation coefficients for most markers, but it estimates more robustly imputation quality for variants with low MAF [67]. Since Pearson's correlation coefficients cannot be calculated in the absence of variation, a marker that appears fixed with the reference in the true set of genotypes, but contains an alternate allele when imputed, cannot have an r computed, but can have an IQS. This idea also applies in all cases when a marker is fixed in the true or imputed set, but not in the other. IQS allows us to identify all of these specific error types, and thus provides a more complete account of imputation accuracy.

In addition to the IQS, the exact nature of each error was catalogued and tallied for each individual and variant. This allowed the errors to be categorized as either false heterozygotes (genotyped AA or BB imputed as AB), false homozygotes (genotyped AB imputed as AA or BB) or completely discordant (BB imputed as AA or vice versa). These more detailed error descriptions, in conjunction with MAF, genome position, and assay-of-origin information, allow for a detailed analysis of how these factors influence imputation accuracy to 850 K in each scenario.

To approximate how well represented each individual was in the composite reference, we created a standardized genomic relationship matrix (GRM) as described in [3] using the GEMMA software [68]. The resulting values provide quantitative measures of how far each individual is diverged from the members of the composite reference panel, i.e. larger values indicate that the individuals are more closely related to the animals in the reference panel. To observe the impact of within-breed genetic similarity on imputation accuracy, we created four breed-specific standardized GRM using test individuals and individuals in the reference with more than 50% Angus (number in test = 50, number in reference = 15,013), Holstein (number in test = 50, number of reference = 5127), Gelbvieh (number of testing = 49, number of reference = 470), and Brahman/Nelore (number of testing = 50, number of reference = 2043) ancestry reported from the CRUMBLER pipeline [63]. Row means were calculated for each individual to quantify the relationship between each test individual and the members of their breed.

Results

The MAF spectrum of the SNPs on the HD and F250 assays for the individuals that composed our reference panel is shown in Figure 1, which also displays data for the SNP50 assay for comparison. The SNP50 and HD assays have similar MAF spectra and include mostly common variants. In addition, the HD assay has an increased density of variants with a MAF ranging from 0.025 to 0.075. However, the F250 assay has a much higher proportion of SNPs with a MAF lower than 0.1, which is more similar to the site frequency spectrum of variants identified from genome resequencing [69].

Imputation accuracy metrics

Numerous statistics have been used to evaluate imputation quality. We compared two widely-used statistics [concordance rate and Pearson's correlation (*r*)] with the imputation quality score (IQS), a metric that has been used in several human studies, but not in livestock [67,70]. We tested each of these metrics on the TAUR dataset at both the level of variants and individuals. For variants, IQS were lower than concordance rates, particularly at lower MAF (Figure 2a). In the TAUR dataset, IQS scores were lower than their corresponding *r* values for 81% of cases (Figure 2b). At moderate to high MAF, these metrics generally agreed with each other. However, when MAF were lower than 0.1, both Pearson's correlations and IQS penalized more heavily the imputation errors made for rare variants and resulted in lower averages and larger variances compared to concordance rates.

Since IQS is a metric for assessing variant accuracy, we used error type/count and Pearson's correlation (*r*) between observed and imputed genotypes to determine the impacts of different intrinsic and extrinsic factors on accuracy of imputation for each individual. Individual *r* values using raw and centered genotypes were highly correlated (Pearson's $r = 0.9993$ and Spearman's $r = 0.9998$). Since these values were so highly correlated, we report only individual correlations calculated from the raw genotype values, hereafter. For our 307 test animals, individual *r* ranged from 0.7466 to 0.9993, but 267 of these individuals had *r* higher than 0.990. In addition to characterizing these metrics, we also identified the type of error (complete discordance, false heterozygote, or false homozygote) that occurred on a SNP and individual basis. Individuals with the lowest *r* values (< 0.85) tended to have significantly more false heterozygote errors than false homozygote errors ($p = 1.475 \times 10^{-5}$), whereas well imputed animals showed no significant difference ($p = 0.7891$).

Comparing multi-breed and within-breed imputation reference panels

We used 50 Gelbvieh animals with both HD and F250 genotypes that were masked to SNP50 genotype density to compare the accuracy of imputation obtained when using a multi-breed composite reference (CR) or a single-breed reference (BR) panel when imputing to 850 K SNPs. Gelbvieh had the most complete genotypes of any open herdbook breed in our reference, making it a best-case scenario for breeds with mixed ancestry. Imputation with the breed-specific imputation panel had a mean IQS score of 0.982 (sd = 0.089). Because the breed-specific panel performed well, overall mean accuracy gains were modest but significant when using the composite panel (IQS mean = 0.990, sd = 0.073, paired T-test $p < 2 \times 10^{-16}$) (Figure 3a and see Additional file 5: Figure S3a). In addition to an increase in mean accuracy, the per-SNP accuracy variance decreased significantly when using the CR compared to the BR reference panel (F-test $p < 2 \times 10^{-16}$). Of the 107,110 SNPs for which IQS changed when imputed against the different reference panels, 89,930 had an increased score with the CR panel (average IQS increase compared to BR = 0.0797), whereas only 15,349 (average IQS decrease compared to BR = 0.0603) had a decreased score. For these two sets of SNPs, the average magnitude of the accuracy increases was significantly greater for the CR panel than for the BR panel ($p < 2 \times 10^{-16}$).

The most substantial accuracy gains from the use of the CR panel were observed for low MAF variants (Figure 3b and see Additional file 5: Figure S3b). Although accuracy gains were modest for variants with MAF higher than 0.1 (0.007 IQS increase), the increase in IQS for rare variants was 0.0182 when imputing with the CR panel. This increase in the quality of low MAF imputation was not detected when using concordance rate or r statistics (Table 3). Of the 122,288 markers that were not perfectly imputed using the BR panel, there was an increase in IQS of 0.059 (r increase 0.032) when imputed with the CR panel.

One concern with using a large multi-breed reference panel for imputation is that it may introduce variation that does not actually exist in the population being imputed. Individuals had significantly fewer false heterozygote errors when using the CR panel compared to the BR panel (paired T-test $p = 0.0039$). There were, on average, 733 fewer false heterozygote calls per individual when the CR panel was used.

Whereas the per-variant increases in imputation accuracy were significant, the most substantial improvements in imputation accuracy due to the use of the CR panel were found for specific individuals. The mean individual r increased significantly from 0.9962 (s.d. = 0.0032) with the BR panel to 0.9979 (s.d. = 0.0010) with the CR panel ($p = 0.0012$). Animals that already had their genotypes accurately imputed using the BR panel did not show significant increases in accuracy with the CR panel. However, animals with the largest number of BR panel-induced imputation errors had much greater increases in accuracy when the CR panel was used (Figure 4). The 14 individuals with more than 5000 total errors when the BR panel was used had, on average, 5522 fewer imputation errors (s.d. = 2361.33) when the CR panel was used for imputation. Conversely, the 35 individuals with less than 5000 imputation errors when the BR panel was used had only 209 fewer imputation errors, on average (s.d. = 609.80), when the CR panel was used for imputation.

Across the MAF spectrum, accuracies for the “one-round” imputation were consistently higher than those for the “two-round” method. However, the overall magnitudes of the differences were modest. The “one-round” imputation increased the overall accuracy of imputation by 0.000762 IQS units, and for low MAF variants by 0.00256 units. In the “one-round” imputation, the addition of imputed rare variants from the F250 into the combined reference also increased the imputation accuracy of rare variants that were exclusive to the HD panel. The HD markers with MAF lower than 0.05 that were imperfectly imputed using the “two-round” method had an average increase in IQS of 0.0846 when imputed by the “one-round” approach (Table 4). For the HD variants with moderate to high MAF, imputation accuracy increased slightly with the “one-round” compared with the “two-round” approach.

Impact of the breed representation in the reference panel on imputation accuracy

Using individual imputation accuracy measures for 307 test animals, we identified the effects of an individual’s breed composition and of those breeds’ representations in the CR panel on individual imputation accuracy. Using the CR panel, individual r ranged from 0.747 to 0.999 while total imputation errors per individual ranged from 932 to 219,737. The accuracy of imputation was strongly related to an animal’s identified breed (Table 5). Individuals from breeds that were adequately represented in the CR panel (Angus, Gelbvieh, Hereford, Holstein, Jersey, Limousin, Nelore and Simmental, Table 2) were generally well imputed (median $r = 0.997$, range = [0.930, 0.999]) (Figure 5). Gelbvieh individuals had the highest mean imputation accuracy ($r = 0.998$), which is likely due to the high proportion of Gelbvieh animals genotyped on both the F250 and HD in the reference panel. Gelbvieh comprised 10.66% of the reference panel individuals with complete 850 K genotypes, second only to Holstein (80.13% of total). Since HD markers represent the largest proportion of the 850 K SNP panel, individuals from breeds with large numbers of HD genotypes, but relatively few F250 genotypes, such as Nelore, were still imputed at high accuracy (median $r = 0.981$, range = [0.9774, 0.9844]).

Individuals from breeds that were only sparsely represented in the CR panel (Brahman, Gir, N'Dama, and Romagnola) had decreased mean accuracies and increased per-animal imputation accuracy variances (mean $r = 0.890$, range = [0.747, 0.961]).

We used a GRM that was created with observed genotypes from all reference and test individuals to determine if an individual's genetic similarity to individuals in the CR panel was related to its imputation accuracy (Figure 6). There was no direct relationship between an individual's average relatedness to members of the CR panel and imputation accuracy. Rather, imputation accuracy was better predicted by the breed representation of the individuals in the CR panel. For example, individuals assigned by the CRUMBLER pipeline as Romagnola had relatively low imputation accuracies (mean individual $r = 0.874$, range = [0.8549, 0.8958]), although their genetic similarity values were comparable to those for the Hereford and Jersey samples. The low imputation accuracy for the Romagnola breed likely stems from the low representation of its haplotypes within the CR panel (15 HD and 8 F250 genotypes). We observed the opposite for Nelore; although the Nelore individuals were distantly related to the members of the CR panel as a whole, the larger number of samples contained in the reference panel (858 HD and 7 F250 genotypes) resulted in accurate imputation (mean individual $r = 0.981$). This was also observed for Gir, which is as diverged from taurines as the Nelore breed, but its reduced imputation accuracy was due to the presence of only 13 HD and nine F250 individuals in the CR panel. The average genetic relationship with individuals of the same breed in the CR panel had varying magnitudes of correlation with individual imputation accuracies, depending on the breed (see Additional file 6: Figure S4). Measures of genetic similarity and individual imputation accuracy were highly correlated in Brahman and Nellore ($r = 0.940$), negatively correlated in Gelbvieh ($r = -0.138$), and moderately correlated in Angus and Holstein ($r = 0.207$ and 0.241, respectively).

Impact of the starting assay number of markers on 850 K imputation

To test the impact of the starting assay number of markers on 850 K imputation accuracy, we used the TAUR dataset masked to represent the contents in markers of five common commercial assays. Each successive increase in assay marker number led to increases in imputation accuracy both overall and for low-MAF variants (Table 6 and Figure 7). The largest increase in imputation accuracy came between the number of markers used in ULD and GGP-LD assays. Imputation accuracies from the ULD were exceptionally poor for low-MAF variants. Although the decline in IQS at low MAF was also observed for other assays, it was much greater for the ULD variants (0.1385 IQS decrease). At marker densities higher than that of the GGP-LD assay, increases in overall imputation accuracy were smaller (GGP-LD → SNP50 = 0.0133, SNP50 → GGP-90KT = 0.0051, and GGP-90KT → GGP-HD = 0.0036). Similar increases in accuracy were observed for low-MAF variants as the starting assay density increased (ULD → GGP-LD = 0.1249, GGP-LD → SNP50 = 0.0152, SNP50 → GGP-90KT = 0.0099, and GGP-90KT → GGP-HD = 0.0099).

Individual accuracies also increased as the starting assay number of markers increased. A one-way ANOVA using Tukey’s method for multiple comparisons indicated a significant difference in 850 K imputation accuracy between the ULD and GGP-LD ($p = 9.05 \times 10^{-5}$) assays, but not between the GGP-LD and SNP50 ($p = 0.1486$) assays (Figure 8). There were no significant differences between the SNP50 and GGP-90KT or GGP-HD assays. However, the starting GGP-LD marker number had a significantly lower imputation accuracy compared to GGP-90KT ($p = 0.0049$). This suggests that imputation accuracy gains are minimal when the starting assay marker number is larger than 50,000 variants (Table 7).

Error profiles and regions of low imputation accuracy

Using the imputation accuracy information for the TAUR dataset, we identified a number of genomic regions for which markers had a low imputation accuracy. Although most markers were accurately imputed, most chromosomes have at least one small region that contained poorly imputed markers (Figure 9). The overall number of poorly imputed markers was quite small. Only 21,848 markers had an IQS lower than 0.8 (1.95% of imputed makers) (see Additional file 5: Figure S3a and S3b), and only 8963 markers had more than 10 imputation errors (1.07% of imputed markers). When using the IQS metric, we found that there are markers imputed with low accuracies on each chromosome, particularly low-MAF variants with relatively few errors (making $IQS = 0$) (Figure 9a). However, both IQS and total error counts (Figure 9b) reveal clusters of markers with a low imputation accuracy. Investigation of these regions indicated that the probe sequences for these variants had multiple equally likely matches to the genome, which indicates either that there were genome mis-assemblies or simply that the wrong location was chosen to represent the position of the marker. The latter can be easily rectified by changing the map files for these variants to reflect the correct alternate position.

Discussion

Imputation accuracy metrics

Most studies on imputation accuracy in livestock populations have used two methods to assess the adequacy of imputation: concordance rate, i.e. the proportion of correctly imputed genotypes, and the Pearson correlation (r) between observed and imputed genotypes. Although both statistics make sense at the level of individuals, their ability to identify markers for which genotypes are poorly imputed is not optimal, particularly for markers with a low MAF. Because our dataset contains a large proportion of rare variants (24.30% markers with MAF < 0.1), a statistic that more robustly represents the quality of imputation is essential. Using the IQS statistic, we show that r and especially concordance rate, overestimate the accuracy of imputation for low-MAF variants [67,71]. In the GEL test data, 2070 variants with an average MAF of 0.040 had high concordance rates (0.97 average), but very low IQS scores (0.0). Unlike r, which requires that markers be variable in both the true and imputed datasets, IQS can be calculated for variants that are not variable in either the observed or imputed datasets. This provides a more complete view of the imputation accuracy at each locus, particularly for those with an extremely low MAF. This information is lost when using r, and imputation accuracies are grossly inflated if measured using concordance rate. That said, r and IQS are highly correlated ($r = 0.9892$) and provide equally useful diagnostics for imputation quality at most sites. We note that while we treated the HD and F250 genotype calls as being correct, a ~ 0.2% error rate is associated with these genotyping platforms [2] (see Additional file 1: Table S4).

Impact of the F250 assay on imputation of rare variants

The F250 assay was designed to query genotypes at a large number of rare, potentially functional variants and is very gene-centric (i.e. they are not evenly spaced). Common variants were also included in the F250 assay design to allow for imputation and genomic prediction applications. The rare variants present on the F250 assay are important in the context of this work for two reasons. First, imputing an additional ~ 170,000 variants at the population level will increase researchers' ability to refine GWAS signals and identify putative QTN due to increased marker density within QTL regions. Second, because the variants are very gene-centric, it is anticipated that the accuracy of imputation to the whole-genome sequence level will be improved within genic regions. The inclusion of rare variants will likely increase the imputation accuracy of other rare variants that are not directly assayed, as strong LD (r^2) requires that allele frequencies at two markers be similar. In the absence of selection, rare variants are assumed to have been recently derived, and thus are likely in LD with other recently derived rare alleles [72]. By adding rare variants to our reference panel with the F250 assay and by genotyping a large number of individuals, we improve the imputation of rare variants that are not directly assayed by the F250. Although many individuals in our reference panel have only imputed F250 genotypes, their presence had a significant impact on the imputation accuracies of rare variants. Whereas at a reduced scale, our comparison of "one-round" vs. "two-round" imputation showed that leveraging rare F250 variants helped impute low-MAF variants that are only assayed by the HD assay (Table 4). We expect that these increases in imputation accuracy of rare variants that are achieved from the use of the F250 assay will be carried over to subsequent imputation to whole-genome sequence-level. The positive impact of the F250 assay on imputation of rare variants underscores the need for additional complete 850 K data in our reference panel (individuals genotyped with both the HD and F250 assays). The highest imputation accuracies were observed for breeds that had the largest numbers of complete 850 K genotypes because more of the haplotypic diversity in those breeds was directly captured in the reference panel.

Multi-breed vs. within-breed imputation reference panels

Early imputation studies primarily concentrated on homogenous populations. When imputation is performed in closely related animals from breeds with small effective population sizes, such as Holstein [73,74], highly accurate imputation can be achieved from using a relatively small set of reference genotypes. Recently, large numbers of genotypes have been produced using low-density assays in outbred animals, admixed individuals, from both registered and commercial populations. In conjunction, many animals from a wide range of breeds have now been genotyped on high-density assays such as HD and F250. By combining all available high-density genotypes into a single multi-breed composite reference panel, we found increased imputation accuracy across the MAF spectrum. Comparing the composite reference panel with a breed-specific reference panel, the most substantial increases occurred at the level of individuals. Genotypes for individuals that were accurately imputed using the breed reference panel saw no substantial increases in accuracy when imputed using the CR panel. However, individuals with poorly imputed genotypes using the BR panel had a substantial reduction in the number of imputation errors when imputed using the CR panel. The increased haplotypic diversity present in the composite reference panel improves the accuracy of imputation of introgressed haplotypes that are not present in a more limited breed-specific reference panel. It is important to be aware that in the context of routine genotyping and imputation, there is no a priori knowledge on which individuals may have poorly imputed genotypes. On the one hand, gains in accuracy from using the CR panel may be small in closed herdbook populations such as Holstein or Angus but they are unlikely to be worse than if the panel is restricted to a breed-specific reference. On the other hand, for open herdbook or composite breeds, increases in imputation accuracy are likely substantial. We did not detect an increase in false heterozygote or false homozygote genotype calls using the multi-breed reference panel, which suggests that the use of a CR panel does not introduce false variation into imputed genotypes at a higher rate than imputation using a within-breed reference panel. For breeds that are adequately represented in the CR panel, we found imputation accuracies (median $r = 0.997$) that were consistent with the error rates of the genotyping assays (see Additional file 1: Table S4), which suggests a near-perfect imputation process.

Previous work recommended the use of multi-breed reference panels for whole-genome sequence imputation [51,75]. Our findings for high-density genotypes with an allele frequency spectrum similar to that of the genome sequence supports this finding and suggests that improvements in imputation accuracy for outbred and admixed populations will benefit from the sequencing and inclusion of diverse animals that will capture more of the haplotypic diversity that is found in cattle. Further improvements in accuracy could be obtained by removing Mendelian inconsistencies from the raw dataset that is used to create the CR panel, which was not performed for this study.

Breed representation in the reference panel

An individual's average relatedness to the entire CR panel was not a good predictor of imputation accuracy. Our multi-breed reference panel was heavily biased towards the most common and economically relevant American beef breeds but also had a diverse array of individuals from other breeds in varying numbers. We found that even low levels of admixture with breeds not adequately represented in the CR panel can lead to decreased imputation accuracies. Information on breed composition was valuable for identifying outlying individuals or breeds that, in theory, should have been accurately imputed. For example, the five individuals labeled as Angus with low imputation accuracies were found to be admixed with breeds that are not well represented in the CR panel (see Additional file 3: Table S6). Each of these individuals identified as Angus actually had relatively low proportions of Angus ancestry (0.107 to 0.532 Angus and Red Angus), and moderately high proportions of breeds sparsely represented in the CR panel. The most significant increases in imputation accuracy will likely come through the addition of high-density genotypes for breeds that are sparsely represented in our reference panel, and through the addition of more completely genotyped individuals, i.e., those with both HD and F250 genotypes. It is worth noting that the breed accuracies reported here for populations with a limited representation in our composite reference panel (Brahman, Gir, N'Dama, Romagnola) would have improved if we had not removed large proportions of each of the breeds to create the test set. We expect that the accuracies reported here are underestimated compared with those achieved by imputation against the full CR panel.

Starting assay marker numbers

The starting assay marker number had a significant impact on the accuracy of imputation to 850 K. In agreement with the conclusions on LD of the Bovine HapMap project, we found that approximately 50 K SNPs are needed to impute to 850 K with high accuracy [76]. This observation likely has a larger impact on research applications that seek to identify QTN rather than applications that are targeted towards genomic prediction. At common allele frequencies (MAF > 0.1), IQS values were steady for all starting assay densities. The decline in imputation quality of rare variants (MAF < 0.1) relative to MAF was much more severe for low-density starting assays, particularly the ULD assay, than for higher density starting assays. When starting array densities are increased above 50 K SNPs, significant gains in imputation accuracy will come almost exclusively from improved imputation at rare variants. There is a large number of individuals that have been genotyped with assays with small numbers of common markers (< 10,000 markers) and these individuals can be accurately imputed to ~ 50 K common markers [73]. Studies that impute from these densities to 850 K and whole-genome sequence should expect significantly more errors. If the aim is to perform both genomic predictions and downstream causal variant discovery, via imputation, our recommendation is to genotype new individuals with an assay density of ~ 50,000 SNPs.

Conclusions

We conclude that, in diverse samples, as seen in typical beef cattle populations, a multi-breed phasing and imputation panel will provide the highest imputation accuracies. Individuals that have a moderately represented ancestry in the reference panel will have genotypes accurately imputed. Imputation accuracies were highest for rare variants when using the composite reference panel. The addition of rare variants from the F250 assay increased the imputation accuracy of rare variants in the HD assay. The addition of a large number of individuals that are genotyped for rare variants will likely improve imputation of rare variants to the sequence level. We confirm that for imputation to 850 K, gains in accuracy reach a plateau as the starting assay marker number exceeds 50 K SNPs. We identified a small subset of SNPs with poor imputation accuracies, most of which seem to be caused by location errors of probe sequences that can be corrected. The largest gains in imputation accuracy are expected to come from the addition of individuals with complete (HD and F250) genotypes, with the largest gains coming from modest increases in the numbers of individuals from the less well-represented breeds. Imputation accuracies for the breeds that are adequately represented in the multi-breed composite-reference panel when the starting assay comprises at least 50 K SNPs should approach accuracies of 1.0 minus the genotyping assay error rate. We anticipate that the CR panel presented here will serve as a foundation reference panel, on which the global cattle community can build to further increase the accuracy of genotype imputation.

Acknowledgements

Holly Neibergs, Jim Reecy, James Koltes, and Raluca Mateescu provided GGP-F250 genotypes for the reference panel. Denis Larkin and Marta Farré provided early access to conserved non-coding elements for the design of the F250 assay.

Authors’ contributions

TNR, JED, and RDS designed the study and drafted the manuscript. TNR designed the imputation pipeline and performed statistical analyses. JFT led the team for designing the GGP-F250 assay and generated GGP-F250 genotypes for 18,271 animals representing 22 breeds or breed types. JLH assisted in early versions of the imputation pipeline. TEC estimated breed ancestry using the CRUMBLER pipeline. All authors read and approved the final manuscript.

Funding

This project was supported by Agriculture and Food Research Initiative Competitive Grant Nos. 2013-68004-20364, 2015-67015-23183 and 2016-68004-24827 from the USDA National Institute of Food and Agriculture.

Availability of data and materials

The datasets analyzed during the current study are not publicly available because the data originated from a variety of sources with different ownership and sharing agreements. However, data are available from the corresponding author on reasonable request. Please contact authors regarding data availability or to receive imputed genotypes from our pipeline.

Competing interests

The University of Missouri receives royalties from the design of the GGP-F250 from

Neogen GeneSeek, and RDS and JFT receive a portion of those royalties. The other authors declare no competing interests.

Figures

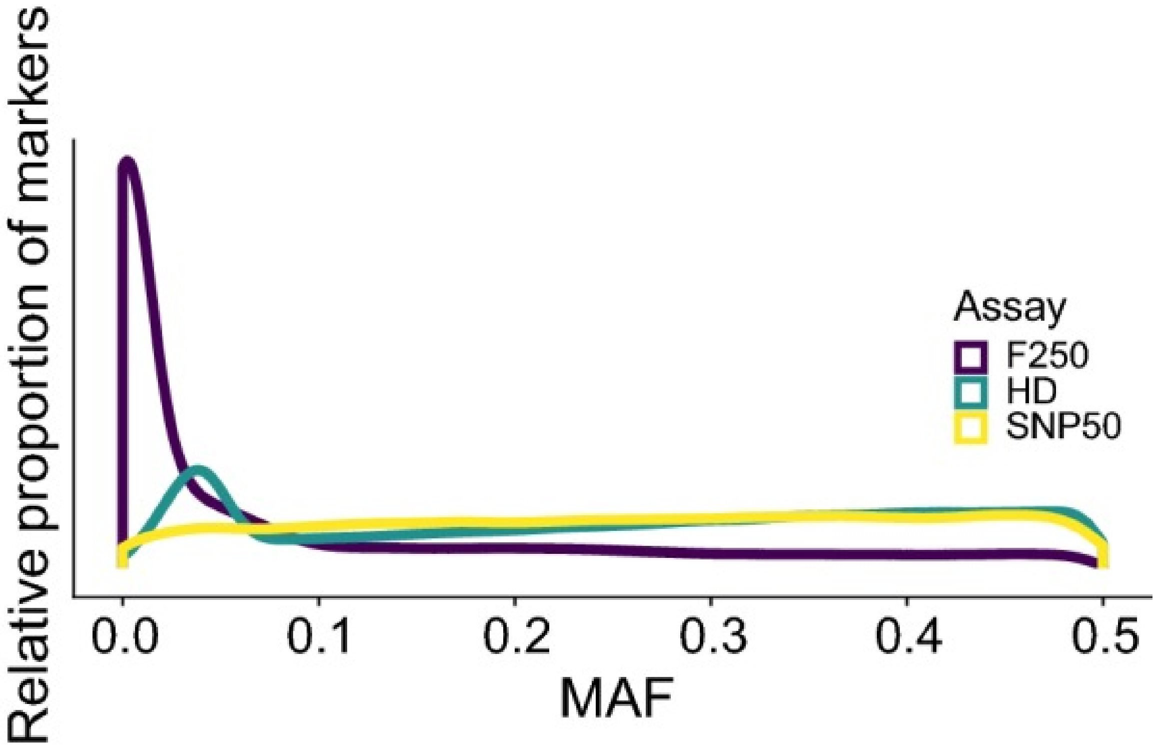


Figure 2.1. Minor allele frequency spectra for three commercially available assays with different marker densities. Density plot of minor allele frequencies for the SNP50 (yellow), F250 (purple), and HD (green) assays

[Figure 1 in text]

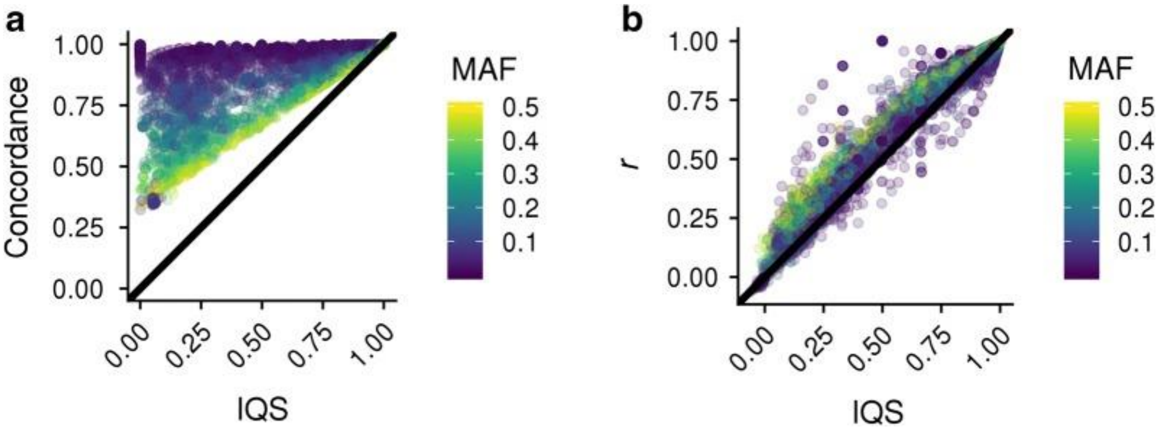


Figure 2.2. The imputation quality statistic (IQS) compared to concordance rate and correlation as measures of imputation accuracy. Three imputation accuracy measures calculated for the TAUR dataset. a concordance, and b Pearson correlation over-estimate imputation accuracies compared to the imputation quality statistic (IQS) resulting in bias and a false high imputation accuracy

[Figure 2 in text]

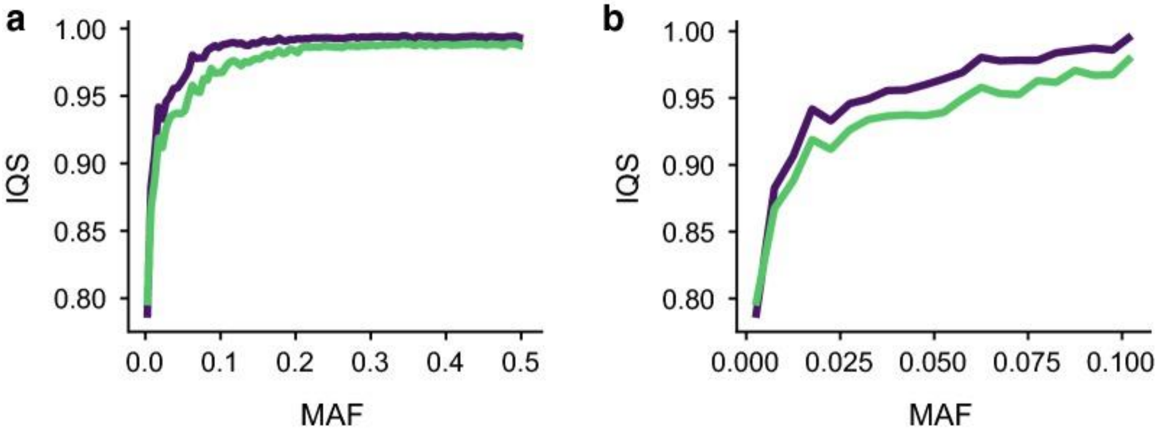


Figure 2.3. The composite reference panel improves per-variant imputation accuracies, particularly for rare variants. Imputation quality statistics when using breed-specific (green) and composite (purple) reference panels for 850 K imputation in the Gelbvieh (GEL) dataset across the MAF spectrum (a), and at low MAF (b) [Figure 3 in text]

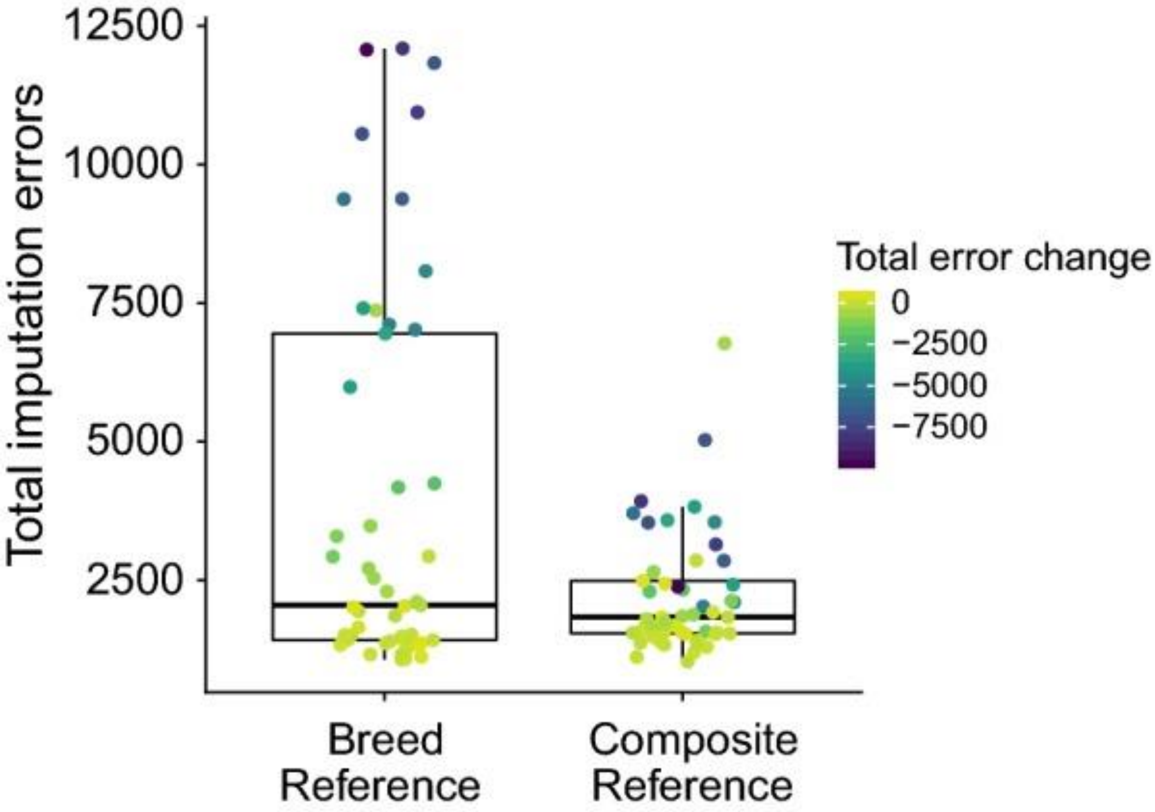


Figure 2.4. Improvements from the composite reference panel were greatest for individuals for which genotypes were poorly imputed. Comparing the total number of errors when imputing from SNP50 to 850 K in the Gelbvieh (GEL) dataset when using breed-specific vs. composite reference panels. Points are individuals, colored by the change in count of errors from the breed to composite reference panels.

[Figure 4 in paper]

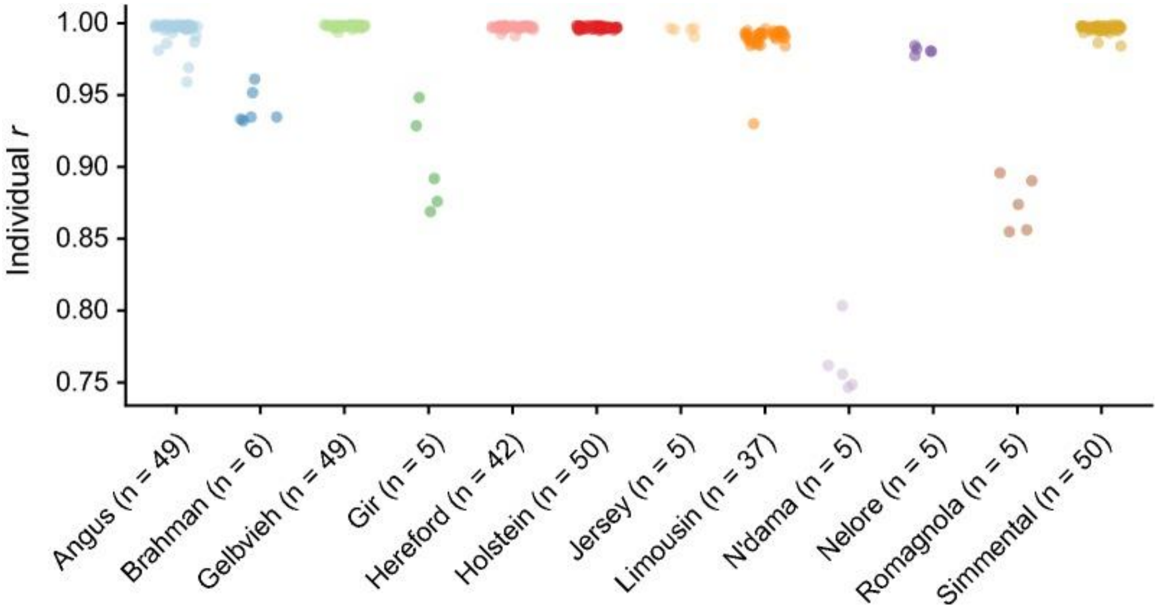


Figure 2.5. Per-individual accuracy by reported breed. Individual *r* by breed. Each point is an individual, colored by breed.

[Figure 5 in text]

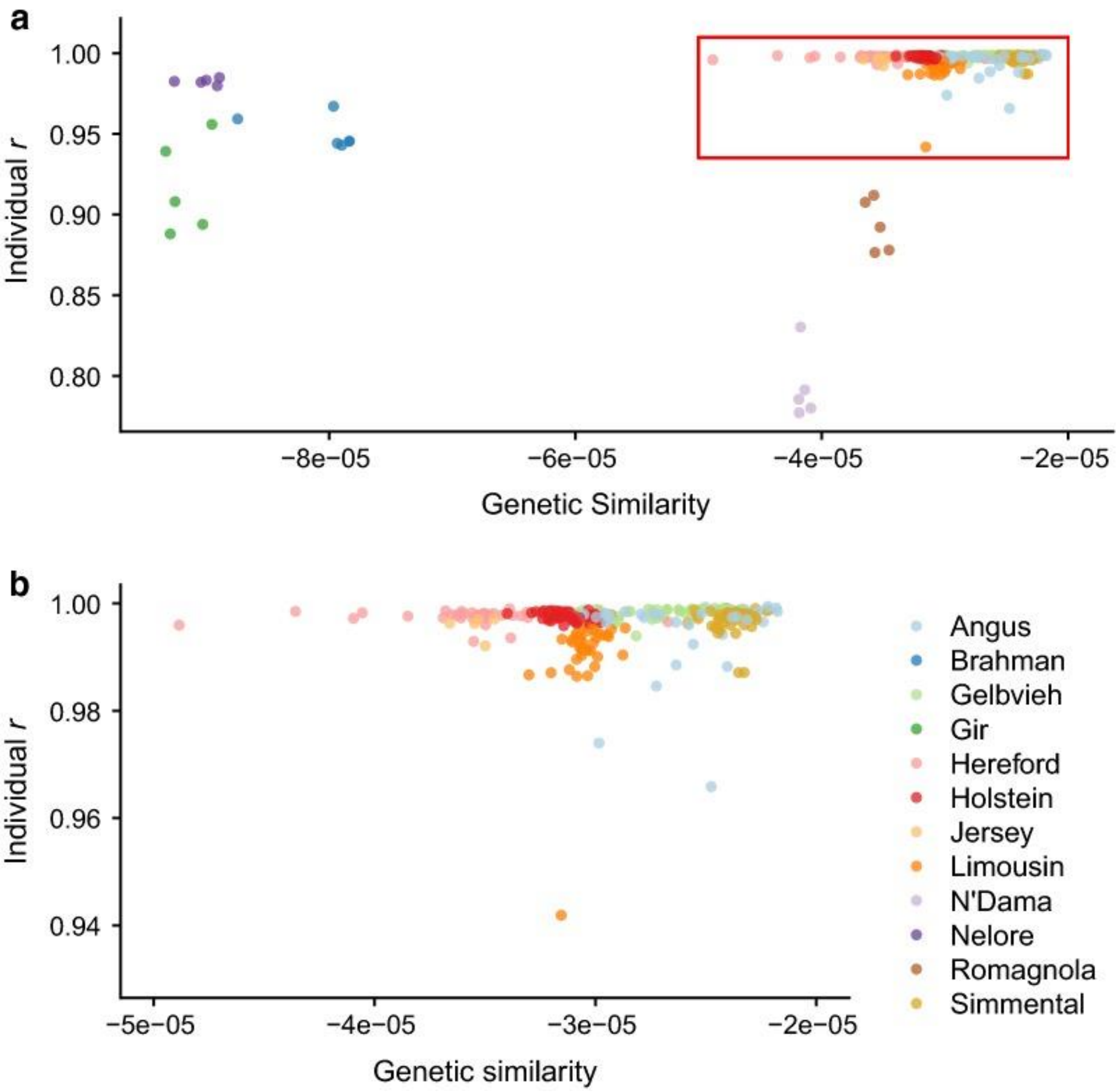


Figure 2.6. Relatedness to composite reference panel members is not a strong predictor of individual imputation accuracy. (a) Per-individual r for the entire testing dataset as a function of individual's genetic similarity to the composite reference. (b) Zoom-in on high-imputation accuracy taurine individuals. Larger values indicate stronger relationships.

[Figure 6 in text]

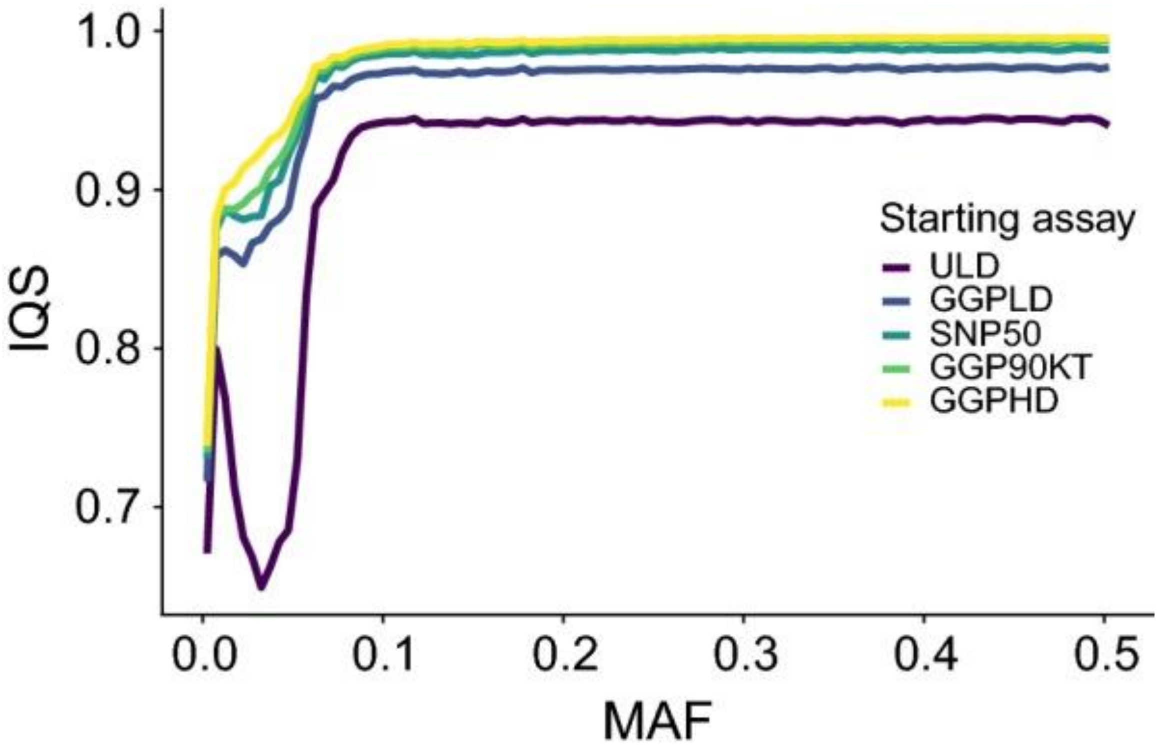


Figure 2.7. Effects of starting assay marker numbers on imputation accuracy across the MAF spectrum. Variant accuracy measures for 850 K imputation in the TAUR dataset based on five assays with different marker numbers. Binned mean IQS lines (per-variant accuracy) across the MAF spectrum.

[Figure 7 in text]

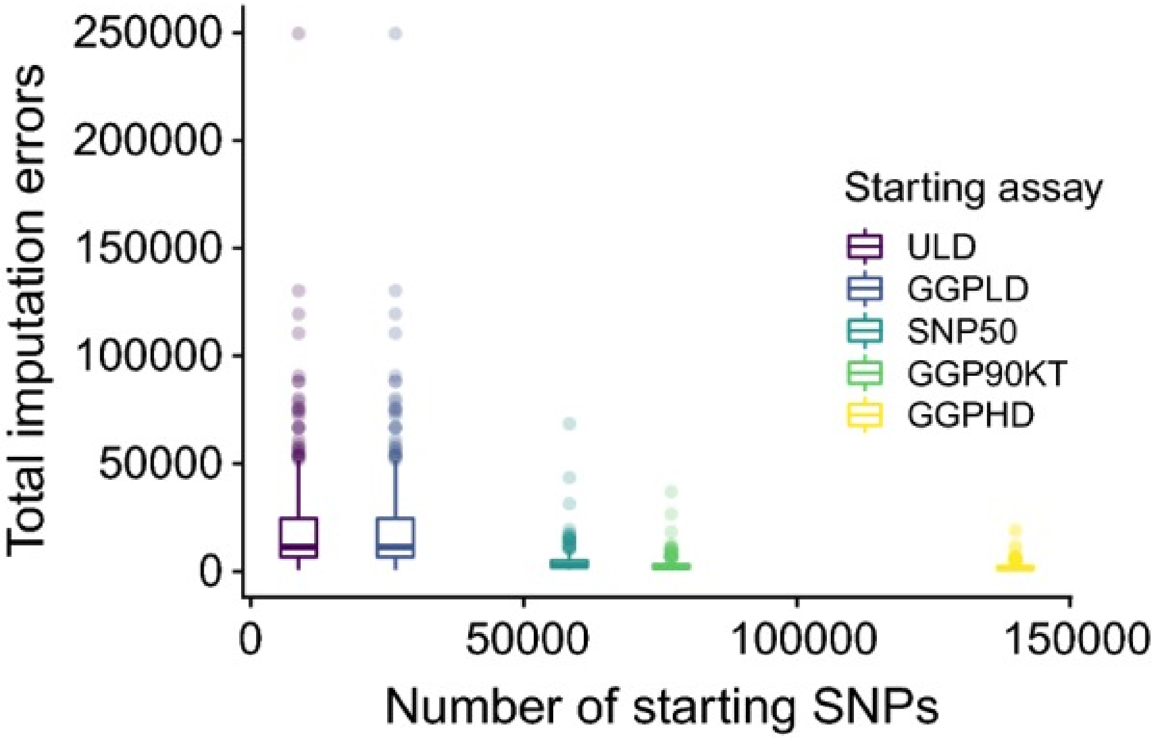


Figure 2.8. Impact of starting assay marker numbers on per-individual imputation accuracy. Per-individual accuracy measures for 850K imputation in the TAUR dataset based on five starting assays differing in marker numbers. Boxplots for total imputation errors based on each starting assay marker number.

[Figure 8 in test]

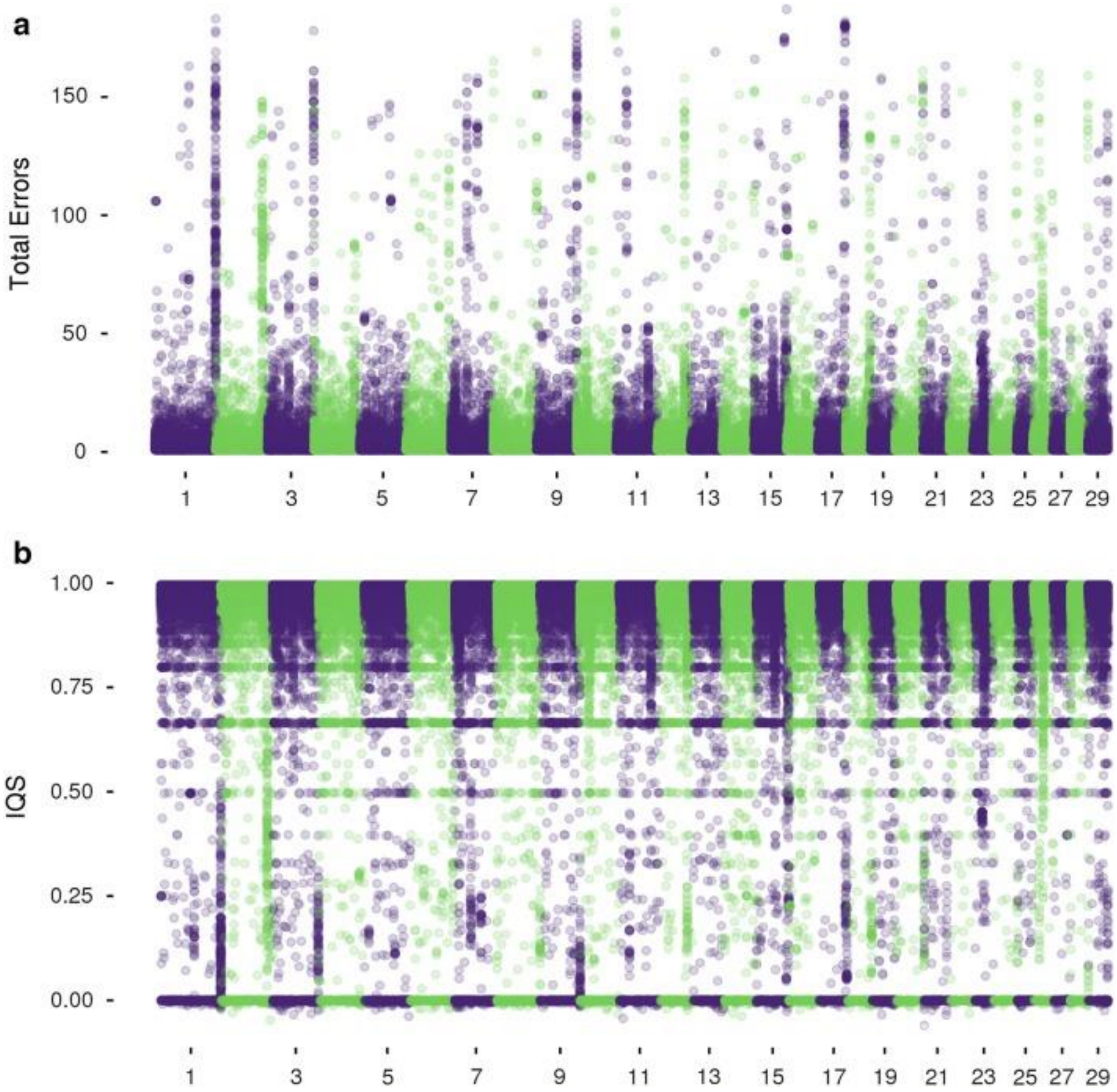


Figure 2.9. Regions with low imputation accuracy exist across the genome but represent only a small subset of the markers. Regions of low imputation accuracy using the TAUR dataset identified by total imputation errors (a), and IQS (b) [Figure 9 in text]

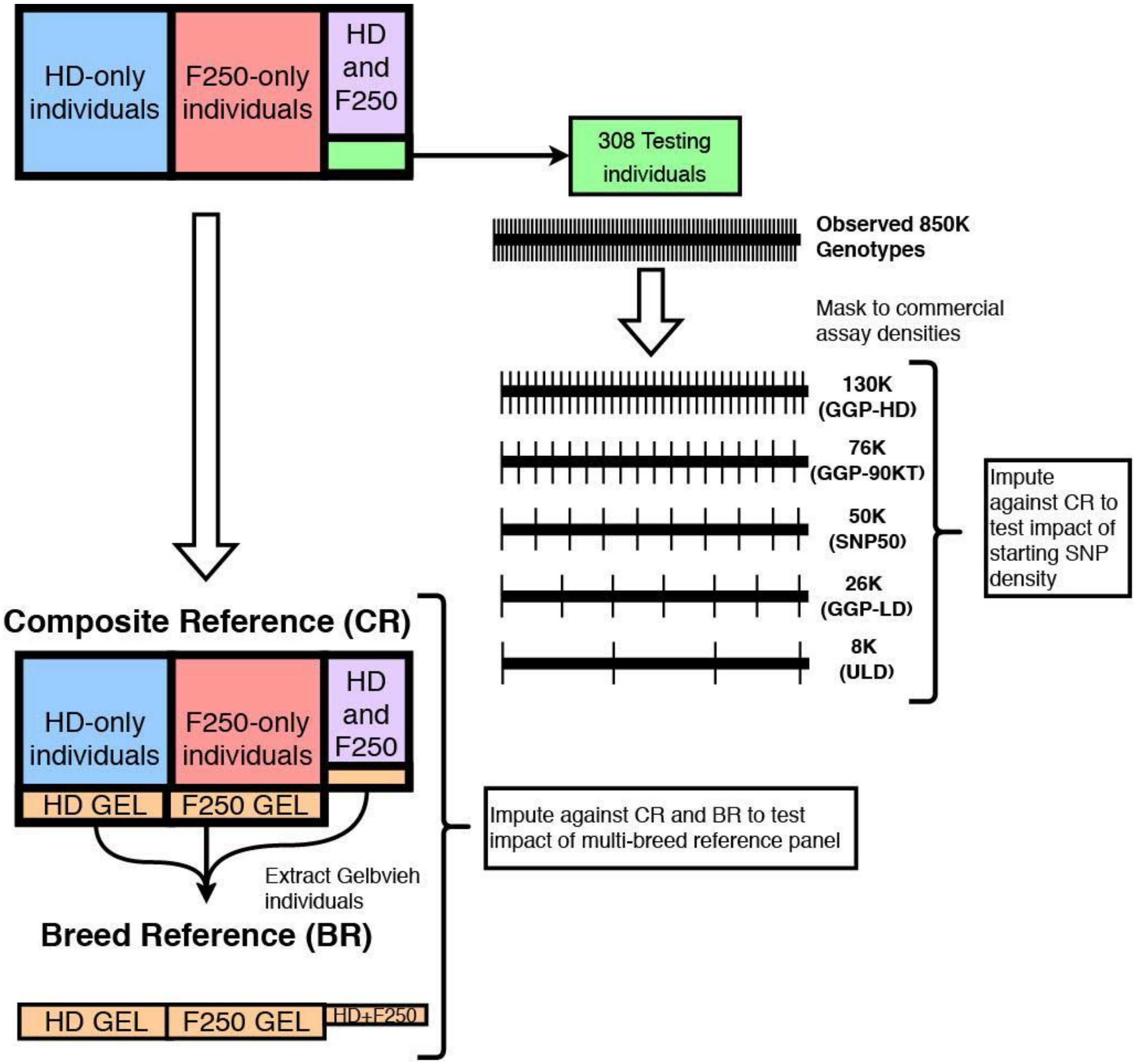
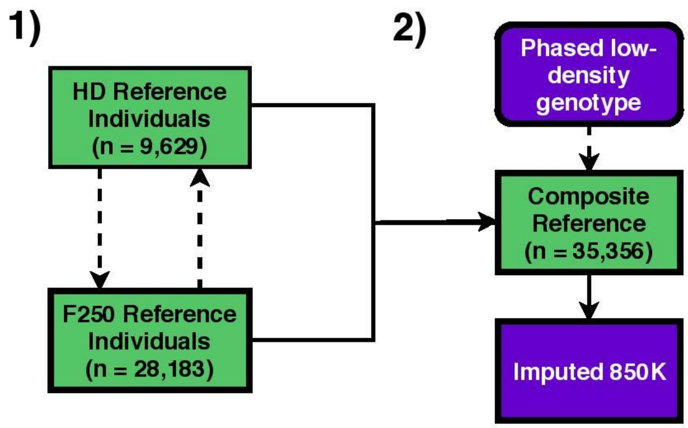


Figure 2.10. Schematic representation of genotype masking for imputation testing

[Figure S1 in text]

One-round imputation


1)
2)
HD Reference Individuals (n = 9,629)
F250 Reference Individuals (n = 28,183)
Phased low-density genotype
Composite Reference (n = 35,356)
Imputed 850K

Two-round imputation

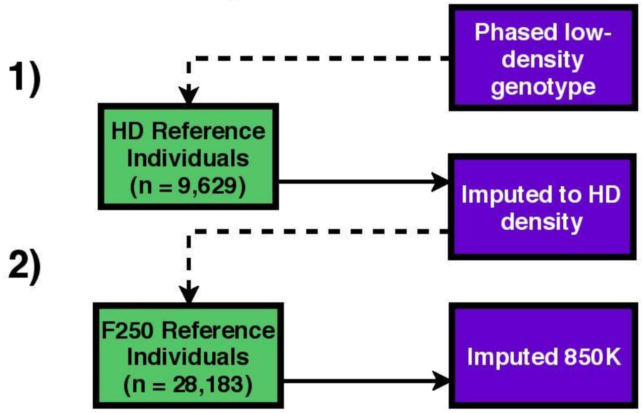

1)
Phased low-density genotype
HD Reference Individuals (n = 9,629)
Imputed to HD density
2)
F250 Reference Individuals (n = 28,183)
Imputed 850K

Figure 2.11. Schematic representation of “one-round” vs. “two-round” imputation.

Description: Dotted lines represent imputation. In “one-round” imputation (a), HD and F250 reference samples are cross-imputed to create a partially imputed composite reference panel (1). This is followed by a single round of imputation of low-density genotypes using the CR panel (2). For “two-round” imputation (b), two rounds of imputation occur: first from low-density to HD (1) and then from HD to 850 K (2)

[Figure S2 in text]

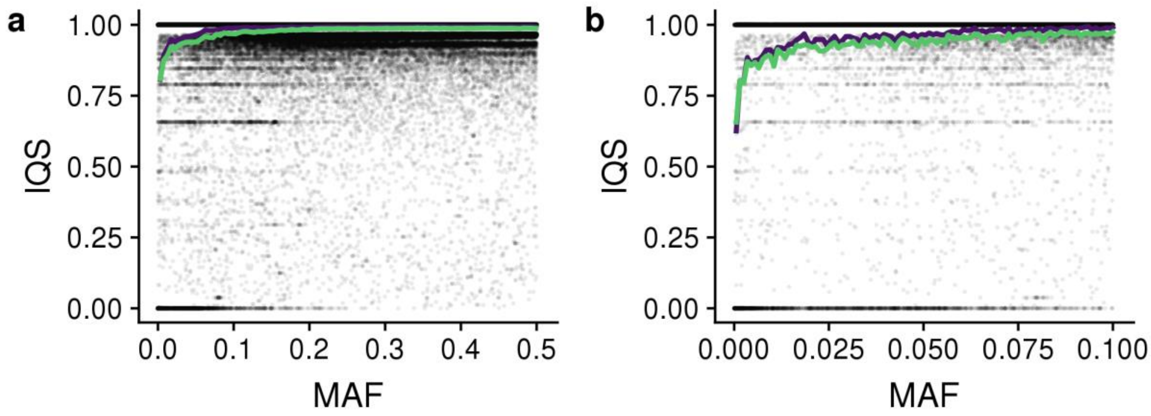


Figure 2.12. Imputation quality metrics when using breed-specific (green) and composite (purple) reference panels for 850 K imputation in the GEL dataset across the entire MAF spectrum (a), and at low MAF (b). Points are individual variants.

[Figure S3 in text]

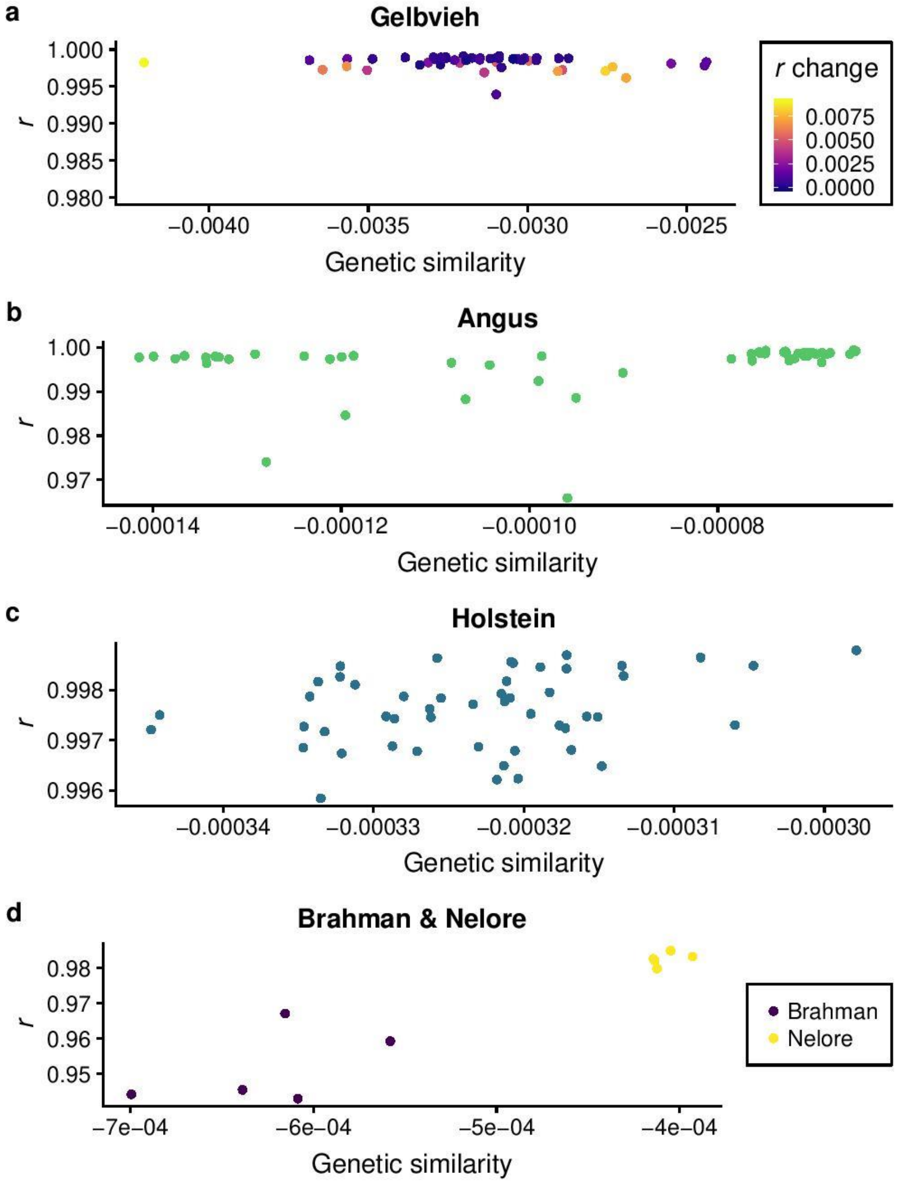


Figure 2.13. Impact of genetic similarity to the reference on imputation accuracy. Genetic similarity is the mean genomic relationship between testing individual and reference individuals with > 50% ancestry of the same breed. Gelbvieh testing individuals (a) are colored by the change in r when using CR versus the BR. (b–d) show r vs. genetic similarity for Angus, Holstein, and Brahman/Nelore respectively.

[Figure S4 in text]

Tables

Table 2.1. Variant counts for masked genotypes of 307 testing individuals used in this analysis before and after filtering

Assay	Starting Assay Density	Filtered Density
GGP-ULD	8,672	6,394
GGP-LDv3	26,504	16,854
SNP50	58,336	44,366
GGP-90KT	76,999	70,581
GGP-HD	139,977	125,446
GGP-F250	227,234	201,236
HD	777,962	753,715

[Table 1 in text]

Table 2.2. Breed representation of composite reference panel after removing 308 test individuals.

Breed	**Number of testing individuals**	**HD and F250[a, b]**	**HD[b, c]**	**F250[b, c]**	**HD and F250[a] (%)**	**HD (%)**
Holstein	50	1932	3170	1944	80.13	32.92
Gelbvieh	49	257	265	514	10.66	2.75
Angus	50	132	2067	14,454	5.47	21.47
Simmental	50	67	427	1759	2.78	4.43
Brahman	5	7	25	632	0.29	0.26
Romagnola	4	4	11	4	0.17	0.11
Nelore	4	3	855	4	0.12	8.88
Jersey	4	3	21	5	0.12	0.22
Gir	5	3	10	6	0.12	0.10
N'Dama	4	3	7	4	0.12	0.07
Brangus	0	0	990	1603	0.00	10.28
Hereford	44	0	569	1834	0.00	5.91
Mixed/crossbred	0	0	419	2830	0.00	4.35
Red Angus	0	0	253	1905	0.00	2.63
Limousin	38	0	215	142	0.00	2.23
Shorthorn	0	0	136	218	0.00	1.41
Charolais	0	0	125	284	0.00	1.30
Santa Gertrudis	0	0	23	11	0.00	0.24
Japanese Black	0	0	19	0	0.00	0.20
Brown Swiss	0	0	15	0	0.00	0.16
Norwegian Red	0	0	5	0	0.00	0.05
Chianina	0	0	2	1	0.00	0.02
Piedmontese	0	0	0	9	0.00	0.00
Braunvieh	0	0	0	7	0.00	0.00
Guernsey	0	0	0	7	0.00	0.00
Beefmaster	0	0	0	3	0.00	0.00
Sheko	0	0	0	2	0.00	0.00
Maine Anjou	0	0	0	1	0.00	0.00

[a] Animals genotyped with both the HD and F250

[b] Number of individuals in CR remaining after 307 testing individuals were removed

[c] Includes individuals genotyped on both HD and F250

[Table 2 in text]

Figure 2.3. Per-variant mean imputation accuracy measures by MAF for Gelbvieh individuals imputed using the breed reference (BR) and composite reference (CR) panels.

MAF bin	BR Concord.[a]	CR Concord.[a]	BR R^2	CR R^2	BR IQS	CR IQS
0.00 - 0.05	0.999	0.999	0.984	0.982	0.910	0.926
0.05 - 0.10	0.997	0.998	0.985	0.99	0.959	0.979
0.10 - 0.15	0.996	0.998	0.986	0.992	0.974	0.989
0.15 - 0.20	0.995	0.997	0.988	0.993	0.982	0.991
0.20 - 0.25	0.994	0.997	0.990	0.995	0.986	0.993
0.25 - 0.30	0.994	0.997	0.990	0.995	0.987	0.993
0.30 - 0.35	0.994	0.997	0.991	0.996	0.988	0.994
0.35 - 0.40	0.993	0.997	0.992	0.996	0.988	0.994
0.40 - 0.45	0.993	0.996	0.992	0.996	0.988	0.994
0.45 - 0.50	0.993	0.996	0.992	0.996	0.988	0.994

[a] Genotype concordance

[Figure 3 in text]

Table 2.4. The mean IQS by MAF for HD-specific markers that were imperfectly imputed using the “two-round” method

MAF Bin	Number of SNPs	Two-round IQS	One-round IQS	IQS Change
0.00 - 0.05	11,788	0.5453	0.6299	0.0425
0.05 - 0.10	18,095	0.9221	0.9330	0.0067
0.10 - 0.15	23,311	0.9724	0.9742	0.0017
0.15 - 0.20	29,222	0.9772	0.9784	0.0012
0.20 - 0.25	34,062	0.9803	0.9812	0.0009
0.25 - 0.30	39,948	0.9807	0.9815	0.0009
0.30 - 0.35	44,309	0.9815	0.9824	0.0009
0.35 - 0.40	47,657	0.9814	0.9822	0.0008
0.40 - 0.45	49,233	0.9815	0.9823	0.0008
0.45 - 0.50	50,908	0.9816	0.9823	0.0007

[Table 4 in text]

Table 2.5. Mean, minimum and maximum individual accuracies (r) by breed for the composite reference 850 K imputation

Breed	Mean Correlation	Min Correlation	Max Correlation
Gelbvieh	0.9979	0.9935	0.9989
Hereford	0.9971	0.9912	0.9988
Holstein	0.9969	0.9947	0.9984
Simmental	0.9963	0.9841	0.999
Angus	0.9953	0.959	0.9993
Jersey	0.995	0.9905	0.9966
Limousin	0.9892	0.93	0.996
Nelore	0.981	0.9774	0.9844
Brahman	0.9412	0.932	0.9611
Gir	0.9027	0.8689	0.9482
Romagnola	0.8742	0.8549	0.8958
N'Dama	0.7632	0.7466	0.8033

[Table 5 in text]

Table 2.6. Per-variant mean and standard deviations for imputation quality statistic (IQS) for 850 K imputation in the TAUR dataset based on the starting assay density

Starting Assay	Starting Density	Mean IQS	SD IQS	Mean IQS (Low MAF)	SD IQS (Low MAF)
ULD	6,394	0.9095	0.1766	0.7720	0.3503
GGPLD	16,854	0.9612	0.1225	0.8969	0.2604
SNP50	44,366	0.9745	0.1154	0.9121	0.2468
GGP90KT	70,581	0.9796	0.1104	0.9220	0.2402
GGPHD	125,446	0.9832	0.1032	0.9319	0.2264

[Table 6 in text]

Table 2.7. Per-individual *r* mean and standard deviation values for 850 K imputation based on starting assay density

Starting Assay	**Starting Density**	**Median R^2**	**SD R^2**
ULD	6,394	0.989	0.6070
GGPLD	16,854	0.995	0.0486
SNP50	44,366	0.997	0.0314
GGP90KT	70,581	0.998	0.0205
GGPHD	125,446	0.999	0.0129

[Table 7 in text]

CHAPTER 3

POWERFUL DETECTION OF POLYGENIC SELECTION AND ENVIRONMENTAL ADAPTATION IN U.S. BEEF CATTLE

Troy N. Rowan[1], Harly J Durbin[1], Christopher M. Seabury[3], Robert D. Schnabel[1,2], Jared E. Decker[1,2]*

[1] Division of Animal Sciences, University of Missouri, Columbia, MO 65211 USA

[2] Institute for Data Science and Informatics, University of Missouri, Columbia, MO 65211 USA

[3] Department of Veterinary Pathobiology, Texas A&M University, College Station, TX 77843 USA

*Corresponding Author

Abstract

Selection on complex traits can rapidly drive evolution, especially in stressful environments. This polygenic selection does not leave intense sweep signatures on the genome, rather many loci experience small allele frequency shifts, resulting in large cumulative phenotypic changes. Directional selection and local adaptation are actively changing populations; but, identifying loci underlying polygenic or environmental selection has been difficult. We use genomic data on tens of thousands of cattle from three populations, distributed over time and landscapes, in linear mixed models with novel dependent variables to map signatures of selection on complex traits and local adaptation. We identify 207 genomic loci associated with an animal's generation number, representing ongoing selection for monogenic and polygenic traits. Additionally, hundreds of additional loci are associated with continuous and discrete environments, providing evidence for local adaptation. These candidate loci highlight the nervous system's central role in local adaptation. While advanced technologies have increased the rate of directional selection in cattle, it has been at the expense of local adaptation, which is especially problematic in changing climates. When applied to large, diverse cattle datasets, these selection mapping methods provide an insight into how selection on complex traits continually shapes the genome. Further, by understanding the genomic loci involved in adaptation, we are able to both breed more adapted and efficient cattle and understand the basis for mammalian adaptation, especially in changing climates.

These selection mapping approaches clarify selective forces and loci in evolutionary, model, and agricultural contexts.

Author Summary

Interest in mapping the impacts of selection and local adaptation on the genome is increasing due to the novel stressors presented by climate change. Until now, approaches have largely focused on mapping "sweeps" on large-effect loci. Highly powered datasets that are both temporally and geographically distributed have not existed. Recently, large numbers of beef cattle have been genotyped across the United States, including influential cryopreserved individuals. This has created multiple powerful datasets distributed over time and landscapes. Here, we map the recent effects of selection and local adaptation in three cattle populations. The results provide insight into the biology of mammalian adaptation and generate useful tools for selecting and breeding better-adapted cattle for a changing environment.

Introduction

As climate changes, organisms either migrate, rapidly adapt, or perish. The genes and alleles that underlie adaptation have been difficult to identify, except for a handful of large-effect variants that underwent selective sweeps [77]. It is becoming increasingly apparent that for adaptation, hard sweeps are likely to be the exception, rather than the rule [78]. Polygenic selection on complex traits can cause a significant change in the mean phenotype while producing only subtle changes in allele frequencies throughout the genome [22]. Most selection mapping methods require discrete grouping of subpopulations, making the identification of selection within a largely panmictic population difficult. Further, in many cases these models are unable to derive additional power from massive increases in sample size [79]. Millions of North American *Bos taurus* beef cattle have been exposed to strong artificial and environmental selection for more than 50 years (~10 generations) [13], making them a powerful model for studying the impacts selection has on genomes over short time periods and across diverse environments.

Though domesticated, beef cattle are exposed to a broad spectrum of unique environments and local selection pressures, as compared to other more intensely managed livestock populations. This suggests that local adaptation and genotype-by-environment interactions play important roles in the expression of complex traits. Understanding genetic interactions with the environment will become increasingly important in changing climates. Herein, we use two methods (**Figure 1**), the first for detecting complex polygenic selection (Generation Proxy Selection Mapping, GPSM), and the second for identifying local adaptation (environmental Genome-Wide Association Studies, envGWAS). Both methods use genome-wide linear mixed models (LMM) incorporating novel dependent variables in large temporally and spatially dispersed datasets, while explicitly controlling for family and population structure as well as uneven sampling (**Figure 1d**). When applied to three US beef cattle populations, each with ~15,000 genotyped individuals, we identified numerous genomic regions harboring directional or environmentally selected mutations. Further, using a meta-analysis approach, we identified loci responding to region-specific selection (**Figure 1e,f**), largely due to the erosion of local adaptation caused by gene flow among ecoregions from the use of artificial insemination sires. This study is the first step in assisting beef cattle producers to identify locally adapted individuals, which will reduce the industry's environmental footprint by increasing efficiency and resilience to stressors. Further, this repurposing of commercially-generated genomic data provides us unprecedented power to gain insight into the biology of adaptation in mammalian species.

Results

Detecting ongoing polygenic selection with Generation Proxy Selection Mapping (GPSM)

Though the first cattle single nucleotide polymorphism (SNP) genotyping assay was developed just over a decade ago [80], numerous influential males who have been deceased for 30 to 40 years have been genotyped from cryopreserved semen (**Figure S1, Table S1**). These powerful datasets provide a temporal distribution of samples spanning at least ten generations for the numerically largest US beef breeds. This temporal distribution of genotypes allows us to search for allelic associations with generation number, or a proxy such as birth date, to identify loci subjected to directional selection [13]. Using a LMM, we tested for associations between an individual's generation proxy (i.e., birth date) and SNP alleles in three US cattle populations using ~830,000 SNPs. We controlled for the confounding effects of population structure, relatedness, and inbreeding by including a random effect accounting for dependency between samples using a genomic relationship matrix (GRM, **Figure 1d**). Significant associations with birth date indicate variants undergoing frequency changes that are greater than expected due to drift over the time span represented in a sampled dataset (**Figure 1a,c**). Our simulations show that GPSM effectively distinguishes between selection and drift under a variety of genetic architectures, selection intensities, effective population sizes, and sampled time periods (**Figure 1, Tables S1-S2**).

We used continuous birth date and high-density SNP genotypes for large samples of animals from three large US beef cattle populations; Red Angus (RAN; n =15,295), Simmental (SIM; n =15,350), and Gelbvieh (GEL; n =12,031) to map loci responding to polygenic selection (**Table S3, Figure S1**). The LMM estimated that the proportion of variance in individuals' birth dates explained by the additive genetic effects of SNPs was large [Proportion of Variance Explained (PVE) = 0.520, 0.588, and 0.459 in RAN, SIM, and GEL, respectively], indicating that we could theoretically predict an animal's birth date from its multi-locus genotypes with an accuracy of ~70%. The ability to predict birth date from genotypic data is created by the time trend in allele frequencies. The PVE estimates indicate that there must be many genome-wide associations between genotype and birth date. The large amount of variance in birth date explained by the SNP genotypes persists even when the analysis is restricted to individuals born in the last 10 years (~2 generations) or 20 years (~4 generations) (**Table S4**), demonstrating that GPSM can leverage the power of our large datasets to detect subtle changes in allele frequency over extremely short periods of time. We removed the link between generation proxy and genotype by randomly permuting the animals' birth date and on reanalysis of the permuted data we observed PVE to decrease to zero (**Table S4**).

The GPSM analyses for these three populations identified 268, 548, and 763 statistically significant SNPs (q-value < 0.1), representing at least 52, 85, and 92 genomic loci associated with birth date in RAN, SIM, and GEL, respectively (**Figure 2a-f, Table S5**). Despite the tendency for genome-wide association studies (GWAS) to be biased in its detection of moderate frequency variants [81], we identify significant associations across the minor allele frequency range in our GPSM simulations and analyses (**Figure 2g-i**). This suggests GPSM can differentiate drift from selection across the allele frequency spectrum. Rapid shifts in allele frequency create highly significant GPSM signals. For example, *rs1762920* on chromosome 28 has undergone large changes in allele frequency in all three populations (**Figure 2g)**, which in turn creates highly significant q-values (2.810×10^{-27}, 2.323×10^{-150}, 2.787×10^{-265} in RAN, SIM, and GEL, respectively). The allele frequency changes observed for this locus are extremely large compared to other significant regions, most of which have only small to moderate changes in allele frequency over the last ~10 generations. When we regressed allele frequency (0, 0.5, or 1.0 representing *AA*, *AB*, and *BB* genotypes per individual) on birth date, the average allele frequency changes per generation (ΔAF) for significant GPSM associations were 0.017, 0.024, and 0.022 for RAN, SIM, and GEL, respectively (**Table S6**). In the analyses of each dataset, GPSM identified significant SNPs with $\Delta AF < 1.1 \times 10^{-4}$. The generally small allele frequency changes detected by GPSM are consistent with the magnitude of allele frequency changes expected for selection on traits with polygenic architectures [22]. Consequently, we suspect that many of these loci would go undetected when using most other selection mapping methods.

We performed a genomic restricted maximum likelihood (REML) analysis to identify how much of the variation in birth date was explained by various classes of GPSM SNPs. We built three GRMs using different SNP sets: One set with GPSM genome-wide significant SNPs ($q < 0.1$), the second with an equivalent number of the next most suggestive GPSM SNPs outside of loci (> 1 Mb from a $q < 0.1$ significant

SNP), and the third an equivalent number of moderate minor allele frequency (MAF) (MAF > 0.15), non-significant SNPs (p > 0.5) intended to represent loci randomly drifting in the population. For each population, we observed that nearly all of the variation in birth date was explained by the significant and suggestive GRMs. While genome-wide significant loci explain the majority of genetic variance associated with birth year, an equivalent number of suggestive, but not significant SNPs have only slightly smaller PVEs (**Table 1**). We suspect that these SNPs are undergoing directional allele frequency changes too small to detect at genome-wide significance, even in this highly-powered dataset. Since GPSM continues to gain power with additional samples, we suspect that future sample size increases will detect more of these signatures of polygenic selection at a genome-wide significance level. Regardless of the number of SNPs used in the drift GRM, the variance associated with drift was consistently minimal (**Table 1**).

As proof-of-concept, GPSM identified known targets of selection. In Simmental, we identified significant associations at three Mendelian loci that explain the major differences in appearance between early imported European Simmental and modern US Simmental (**Figure 2h)**. These loci: *POLLED* (absence of horns [43]), *ERBB3/PMEL* (European Simmental cream color [82]), and *KIT* (piebald coat coloration [83]) have not appreciably changed in allele frequency since 1995, making their GPSM signature significant, but less so than other loci actively changing in frequency.

In addition to these three known Mendelian loci, we detected numerous novel targets of selection within and across the populations. While the majority of the genomic regions detected as being under selection were population-specific (79.8%, 79.8%, and

77.2% of the significant regions in RAN, SIM, and GEL, respectively), we identified seven loci that are under selection in all three populations, and fifteen more under selection in two (**Table S7**). While GPSM is able to detect Mendelian selection, the overwhelming majority of signatures identified represent selection on complex, quantitative traits. Of the regions identified in multiple populations, many correspond to genes with predicted production-related functions in cattle (*DACH1*-Growth, *LRP12*-Growth, *MYBPH*-Muscle Growth, *RHOU*-Carcass Weight, *BIRC5*-Feed Intake). However, GPSM did not identify any of the well-established large-effect growth loci (i.e., *PLAG1, LCORL*). Growth phenotypes (e.g., birth, weaning, and yearling weights) are known to be under strong selection in all three populations [84], but antagonistic pleiotropic effects such as increased calving difficulty prevent directional selection from changing frequencies at these large-effect loci. We also identified immune function genes under polygenic selection (*ARHGAP15, ADORA1, CSF2RA*). While immune function has not been directly artificially selected in cattle, healthy cattle perform better than their sick counterparts [85]. We also identify a GPSM signature near *PRDM9* on chromosome 1 in all three populations. This paralog has been previously-implicated as a target of selection in both cattle and other mammalian species for its role in modulating recombination [86,87]. Ongoing recent selection on *PRDM9* may indicate either selection for differential recombination rates, or enhanced binding to novel motifs throughout the genome. Many of the selection signatures that were identified in at least two of the populations have no known functions or phenotype associations in cattle, highlighting the ability of GPSM to identify novel, important loci under polygenic selection without the need for any phenotype data.

Biological processes and pathways enriched in genes located proximal to GPSM SNP associations point to selection on drivers of production efficiency and on population-specific characteristics (**Table S8**). In each population, we identified numerous biological processes involved in cell cycle control, which are directly involved in determining muscle growth rate [88], as being under selection. In Red Angus and Gelbvieh we identified multiple cancer pathways as being under selection. This likely represents further evidence of selection on cell cycle regulation and growth rather than on any cancer related phenotypes [89]. Red Angus cattle are known to be highly fertile with exceptional maternal characteristics [90]. We identified the “ovarian steroidogenesis” pathway as being under selection, a known contributor to cow fertility [91]. We also identify numerous other processes involved in the production and metabolism of hormones. Hormone metabolism is a central regulator of growth in cattle [92], but could also represent selection for increased female fertility in Red Angus. Further, Tissue Set Enrichment Analyses (TSEA) of Red Angus GPSM candidate genes showed suggestive expression differences ($p < 0.1$) in multiple human reproductive tissues (**Tables S9-S10**). Enrichments in these tissues did not exist in TSEA of Simmental or Gelbvieh GPSM gene sets, suggesting explicit within-population selection on fertility. Gelbvieh cattle are known for their rapid growth rate and carcass yield. Selection on these phenotypes likely drives the identification of the six biological processes identified which relate to muscle development and function in the Gelbvieh GPSM gene set. Consequently, this gene set is significantly enriched for expression in human skeletal muscle (**Tables S9-S10**). A complete list of genomic regions under population-specific selection and their associated candidate genes is in **Table S5**.

Detecting environmental adaptation using envGWAS

Using an equivalent form of model to GPSM, but with continuous environmental variables (30 year normals for temperature, precipitation, and elevation) or statistically-derived discrete ecoregions as the dependent variable (rather than birth year in GPSM) allows us to identify environmental adaptive loci that have been subjected to artificial and, perhaps in this context more importantly, natural selection [93]. We refer to this method as environmental GWAS (envGWAS). envGWAS extends the theory of the Bayenv approach of Coop et al. (2010) which searches for allele frequency correlations along environmental gradients to identify potentially adaptive loci [94]. Our approach is similar to that used in Yoder et al. 2014, but applied to panmictic, biobank-sized mammalian populations [95]. Unlike many genome-environment association analyses which only used linear models [96,97], our large dataset and the use of multivariate models provides power to identify association while importantly controlling for geographic dependence between samples using a genomic relationship matrix (**Figure S2, Figure S6)**. We used *K*-means clustering with 30-year normal values for temperature, precipitation, and elevation to partition the United States into 9 discrete ecoregions (**Figure 3a**). These ecoregions are largely consistent with those represented in previously-published maps from the environmetrics and atmospheric science literature [98], and reflect well-known differences in cattle production environments. The resulting ecoregions capture not only combinations of climate and environmental variables, but associated differences in forage type, local pathogens, and ecoregion-wide management differences to which animals are exposed. Thus, using these ecoregions as case-control phenotypes in envGWAS allowed us to detect more complex environmental associations. The three studied populations are not universally present in all ecoregions (**Figure 3b, Figure S3b & S4b, Table S11**) and since the development of these US populations in the late 1960s and early 1970s, registered seedstock animals from these populations have a small footprint in desert regions with extreme temperatures and low rainfall.

Although environmental variables and ecoregions are not inherited, the estimated PVE measures the extent to which genome-wide genotypes change in frequency across the environments in which the animals were born and lived. The PVE explained by SNPs ranged from 0.586 to 0.691 for temperature, 0.526 to 0.677 for precipitation, and 0.585 to 0.644 for elevation (**Table S12**). In Red Angus, PVE for ecoregion membership ranged from 0.463 for the Arid Prairie to 0.673 for the Fescue Belt (**Table S13**). We observe similar environmental PVE in both Simmental and Gelbvieh datasets. These measures suggest that genetic associations exist along both continuous environmental gradients and within discrete ecoregions. Despite this genetic signal, principal component analysis (PCA) does not suggest that ecoregion-driven population structure exists in any of the populations (**Figure S5**). Permutation tests that shuffled environmental dependent variables, removing the relationship between the environment and the animal's genotype, resulted in all PVEs being reduced to ~ 0, strongly suggesting that the detected associations between genotype and environment were not spurious. An additional permutation test that permuted animals' zip codes, such that all animals from a given zip code were assigned the same "new" zip code from a potentially different ecoregion provided similar results, indicating that bias due to sampling at certain zip codes was not producing envGWAS signals. From 10 rounds of permutation, there were no SNP associations with p-values < 1×10^{-5}. Consequently, we used this empirically-derived p-value threshold to determine SNP significance in all of the envGWAS analyses, which is also in agreement with the significance threshold used by the Wellcome Trust Case Control Consortium [99]. Gene drop simulations suggest that a portion of the identified associations are likely due to pedigree structure or founder effects (Supplementary Text). However, in this data, the pedigree structure reflects selection decisions of farmers and ranchers that are not beyond the influence of performance differences relative to environmental differences.

Discrete ecoregion envGWAS

In Red Angus, we identified 54 variants defining 18 genomic loci significantly associated with membership of an ecoregion in the discrete multivariate envGWAS analysis (**Figure 3c**). Of these loci, only two overlapped with loci identified in the continuous envGWAS analyses, suggesting that using alternative definitions of environment in envGWAS may detect different sources of adaptation. Of the 18 significant loci, 17 were within or near (< 100 kb) candidate genes (**Tables S14-S15**), many of which have potentially adaptive functions. For example, envGWAS identified SNPs immediately (22.13 kb) upstream of *CUX1* (Cut Like Homeobox 1) gene on chromosome 25. *CUX1* controls hair coat phenotypes in mice [100]. Alleles within *CUX1* can be used to differentiate between breeds of goats raised for meat versus those raised for fiber [101]. The role of *CUX1* in hair coat phenotypes makes it a strong adaptive candidate in environments where animals are exposed to heat, cold, or toxic ergot alkaloids from fescue stress [102]. Other candidate genes identified by envGWAS have previously been identified as targets of selection between breeds of cattle (*MAGI2*, *CENPP*), or in other species (*DIRC1*-humans, *GORASP2*-fish, *ADRB1*-dogs) (**Table S14**). Adaptive signatures shared between cattle and other species may point to shared biological processes that drive environmental adaptation. We also identified four adaptive candidate genes known to possess immune functions (*RASGEF1B*, *SPN*, *ZMYND8*, *LOC100298064*/*HAVCR1*). The envGWAS identified variants within or near immune function genes under ongoing selection in all three populations, thereby suggesting that genetic adaptations conferring resistance or tolerance to local pathogens and immune stressors may be as important as adaptations to abiotic stressors like heat or cold stress.

In Simmental, we identified 11 loci tagged by 39 variants significantly associated with membership of an ecoregion in the multivariate envGWAS analysis (**Figure S3**). In Gelbvieh, 66 variants identified 33 local adaptation loci (**Figure S4**). In the analyses of all three datasets, we identified a common local adaptation signature on chromosome 23 (peak SNP *rs1023574*). Multivariate analyses in all three populations identified alleles at this SNP to be significantly associated with one or more ecoregions (q = 1.24×10^{-13}, 3.15×10^{-12}, 4.82×10^{-5} in RAN, SIM, and GEL, respectively). In all three datasets, we identified *rs1023574* as a univariate envGWAS association with membership of the Forested Mountains ecoregion. However, the most significant univariate association in Red Angus was with the Arid Prairie region which was excluded from both the Simmental and Gelbvieh analyses due to low within-region sample size. In the multivariate analysis for Red Angus, the associated locus spanned 18 SNPs from (1,708,914 to 1,780,836 bp) and contained the pseudogene *LOC782044*. The nearest annotated gene, *KHDRBS2* (KH RNA Binding Domain Containing, Signal Transduction

Associated 2) has previously been identified by other adaptation studies in cattle, sheep, and pigs [103–105]. This variant was not significantly associated with any continuous environmental variable in Red Angus. However, *rs1023574* was significantly associated with temperature, elevation, and humidity variables in Simmental. The *KHDRBS2* locus was preferentially introgressed between *Bos taurus* and domestic yak [106]. Further, this locus shows an abnormal allele frequency trajectory (**Figure 4c**), indicating that it may be a target of balancing selection.

Continuous environmental variable envGWAS

Using continuous temperature, precipitation, and elevation data as quantitative dependent variables in a multivariate envGWAS analysis of Red Angus animals, we identified 46 significantly associated SNPs (**Figure 3g**). These SNPs tag 17 loci, many of which are within 100 kb of strong candidate genes. Univariate envGWAS identified 23, 17, and 10 variants associated with temperature, precipitation, and elevation, respectively (**Figure S6**). The most significant multivariate association in Red Angus is located on chromosome 29 within *BBS1* (Bardet-Biedl syndrome 1), which is involved in energy homeostasis [107]. *BBS1* mutant knock-in mice show irregularities in photoreceptors and olfactory sensory cilia [108] functions that are likely important to an individual's ability to sense its local environment. This region was not significantly associated in any of the univariate analyses of environmental variables, and was not identified in any of the discrete ecoregion envGWAS. Of the candidate genes identified in this Red Angus analysis, 9 have previously been implicated in adaptive functions in humans or cattle (*DIRC1*, *ABCB1*, *TBC1D1*, *AP5M1*, *GRIA4*, *LRRC4C*, *RBMS*3, *GADL1*, *ADCYAP1*, *CUX1,* and *PLA2G12B*) (**Table S16**). Significant SNPs and their corresponding candidate genes for all three datasets are reported in **Table S15**.

While we found few candidate genes to overlap between populations, we identified multiple shared biological pathways and processes (**Table S17**) derived from lists of envGWAS candidate genes. Pathways in common between populations were driven by largely different gene sets. Across all populations, we identified the "axon guidance" pathway, and numerous gene ontology (GO) terms related to axon development and guidance as under region-specific selection. Ai et al. (2015) suggested that axon development and migration in the central nervous system is essential for the maintenance of homeostatic temperatures by modulating heat loss or production [109]. Further, the direction and organization of axons is an essential component of the olfactory system which is frequently implicated in environmental adaptation through the recognition of local environmental cues [17]. In addition to axonal development, a host of other neural signaling pathways were identified in multiple populations. A genome-wide association study for gene-by-environment interactions with production traits in Simmental cattle by Braz et al. (2020) identified a similar set of enriched pathways [33]. These common neural signaling pathways identified by envGWAS are regulators of stress response, temperature homeostasis, and vasoconstriction [110]. We identified other shared pathways involved in the control of vasodilation and vasoconstriction (relaxin signaling, renin secretion, and insulin secretion). Vasodilation and vasoconstriction are essential to physiological temperature control in cattle and other species [111]. The ability to mount a physiological response to temperature stress has a direct impact on cattle performance, making vasodilation a prime candidate for environment-specific selection. Further, vasodilation and vasoconstriction likely also represent adaptation to hypoxic, high elevation environments. Pathways and processes identified by envGWAS signals are reported in **Table S17.**

To further explore the biology underlying adaptive signatures, we performed Tissue Set Enrichment Analysis of our envGWAS candidate gene lists. These analyses, using expression data from humans and worms (*C. elegans*), identified brain and nerve tissues as the lone tissues where envGWAS candidate genes show significantly enriched expression (**Tables S18-S21**). Tissue-specific expression in the brain further supports our observed enrichment of local adaptation pathways involved in neural signaling and development.

Identifying loci undergoing region-specific selection with GPSM ecoregion meta-analysis

envGWAS detects allelic associations with continuous and discrete environmental variables, but does not address whether selection is towards increased local adaptation, or whether local adaptation is being eroded by the exchange of germplasm between ecoregions via artificial insemination. We used the spatiotemporal stratification of genotyped animals to identify loci undergoing ecoregion-specific selection. We performed GPSM within each sufficiently genotyped ecoregion and identified variants with high effect size heterogeneity (Cochran's Q statistic) between ecoregions. Variants with significant heterogeneity across regions that were also significant in at least one within-region GPSM analysis imply ecoregion-specific allele frequency change. These changes could have been due either to selection for local adaptation (**Figure 1e**), or locally different allele frequencies moving towards the population mean (**Figure 1f**). We identified 59, 38, and 46 significant SNPs in Red Angus, Simmental, and Gelbvieh, respectively undergoing ecoregion-specific selection. These represent 15, 21, and 26 genomic loci (> 1 Mb to nearest next significant SNP) (**Figure 4a**). In most cases, these variants have an effect (posterior probability of an effect: m-value > 0.9) in only one or two ecoregions (**Figure 4b**). Further, nearly all represent the decay of ecoregion-specific allele frequencies towards the population mean (**Figure 4c**) as opposed to on-going directional selection for ecoregion specific beneficial adaptations (**Figure S9-S11**).

Despite the apparent ongoing decay of local adaptation, this meta-analysis of ecoregion-specific GPSM identified several interesting candidate genes for environmental adaptation. A significant locus on chromosome 1 at ~73.8 Mb (lead SNP *rs254372*) lies within *OPA1* (OPA1 mitochondrial dynamin like GTPase), which was under selection in the Fescue Belt ecoregion. *OPA1* has been implicated in the regulation of circadian rhythm in mice [112], and is known to regulate metabolic and cardiac adaptations [113] through mitochondrial interactions. We also identified variants within *ADAMTS16* (ADAM Metallopeptidase With Thrombospondin Type 1 Motif 16), which regulates blood pressure in mice [114] and has previously been identified in other adaptation studies [115]. These genes are of particular interest, because the primary symptoms of fescue toxicosis are due to vasoconstriction caused by ergot alkaloids synthesized by endophytes in fescue [28]. Adaptive alleles at these loci are being driven in frequency towards the population mean allele frequency (**Figure 4c**), which is typically a low minor allele frequency.

Discussion

We leveraged large commercially-generated genomic datasets from three major US beef cattle populations to map polygenic selection and environmental adaptation using novel GWAS applications [50]. Using temporally-stratified genotype data we detected very small selection-driven changes in allele frequency throughout the genome. This is consistent with expectations of polygenic selection acting on a large number of variants with individual small effects. Which phenotypes are being selected and driving the allele frequency changes at particular loci is not definitively known. GPSM is a heuristic model, and as a result the SNP effects are not immediately intuitive to interpret in a population genetic context. That said, it allows us to identify the genomic loci responding to selection, and particularly subtle changes due to polygenic selection. GPSM is agnostic to the selected phenotypes, and identifies important loci changing in frequency due to selection without the need to measure potentially difficult or expensive phenotypes. Further, GPSM differentiates between selection and drift while accounting for confounding effects such as uneven generation sampling, population structure, relatedness, and inbreeding. With the availability of large samples our analytical frameworks solve the long-standing population genetics problem of identifying the loci subjected to polygenic selection.

Future studies exploring the effects of selection from the context of complex trait networks could explain how hundreds or thousands of selected genes act together to shape genomic diversity under directional selection. Candidate genes identified by GPSM identify pathways and processes involved in production efficiency (growth, digestion, muscle development, and fertility). In addition to a small number of loci, for which function is known, we identify hundreds of novel signatures of ongoing selection.

The envGWAS identified 174, 125, and 130 SNPs associated with both continuous or discrete environmental factors in Red Angus, Simmental, and Gelbvieh, respectively. Identified candidate genes have functions related to environmental adaptation. Using these environmentally-associated candidate genes we identified an enrichment of pathways and tissues involved in neural development and signaling. These envGWAS associations emphasize the role that the nervous system plays in recognizing and responding to environmental stress in mammals, which will be valuable as society and agriculture cope with climate change. In addition to neural pathways, we observe significantly enriched expression of envGWAS genes in the brain tissues of humans, mice, and worms. Other pathways associated with environmental adaptation reveal the importance of mechanisms involved in regulating vasoconstriction and vasodilation, both of which are essential for responses to heat, cold, altitude, and toxic fescue stressors in cattle.

The statistical power and wide geographical distribution of the cattle comprising these data highlights that the utilized approaches can be leveraged to understand the genomic basis of adaptation in many other studies and species. The small allele frequency differences identified by envGWAS are consistent with a polygenic model of local adaptation, likely driven by small changes in gene expression [116]. Further, envGWAS identifies candidate genes (i.e. *KHDRBS2*) and pathways previously implicated as domestication-related [106]. This suggests that these genes are under natural and balancing selection to cope with environmental stress, and not specifically part of the domestication process. Further, because different genes in the same pathways were detected in the analyses of the different populations, we hypothesize that these pathways influence local adaptation in many mammals and should be studied in other ecological systems. This knowledge will become increasingly valuable as species attempt to adjust to a changing climate.

Artificial insemination in cattle has allowed the ubiquitous use of males which have been found to be superior when progeny performance has been averaged across US environments. Our results suggest that environmental associations are widespread in cattle populations, but that the widespread use of artificial insemination has caused US cattle populations to lose ecoregion-specific adaptive variants. We identified 16, 21, and 30 loci undergoing ecoregion-specific selection in Red Angus, Simmental, and Gelbvieh, respectively. In almost every case, selection has driven allele frequencies within an ecoregion back towards the population mean allele frequency (**Figure 1F and Figure 4C**). In three independent datasets, we identified a single shared environmentally-associated locus near the gene *KHDRBS2*. This locus has been identified as introgressed in yak, and exhibits an irregular allele frequency trajectory which suggests that it may be subject to balancing selection [117]. Though we identified only a single common envGWAS locus, we observed significant overlap in the pathways regulated by candidate genes within the associated loci. This reveals that adaptive networks are complex and that adaptation can be influenced by selection on functional variants within combinations of genes from these networks. As we work to breed more environmentally-adapted cattle, there will be a need for selection tools that incorporate genotype-by-environment interactions to ensure that cattle become increasingly locally adapted.

We demonstrate that large commercially-generated genomic datasets from domesticated populations can be leveraged to detect polygenic selection and local adaptation signatures. The identification of adaptive loci can assist in selecting and breeding better adapted cattle for a changing climate. Further, both our statistical approaches and biological findings can serve as a blueprint for studying complex selection and adaptation in other agricultural or wild species. Our results suggest that neural signaling and development are essential components of mammalian adaptation, meriting further functional genomic study. Finally, we observe that local adaptation is declining in cattle populations, which will need to be preserved to sustainably produce protein in changing climates.

Materials and Methods

Genotype Data:

SNP assays for three populations of genotyped *Bos taurus* beef cattle ranging in density from ~25K SNPs to ~770K SNPs were imputed to a common set of 830K SNPs using the large multi-breed imputation reference panel described by Rowan et al. 2019 [7]. Genomic coordinates for each SNP were from the ARS-UCD1.2 reference genome [62]. Genotype filtering for quality control was performed in PLINK (v1.9) [61], reference-based phasing was performed with Eagle (v2.4) [64], and imputation with Minimac3 (v2.0.1) [66]. Following imputation, all three datasets contained 836,118 autosomal SNP variants. All downstream analyses used only variants with minor allele frequencies > 0.01.

Upon filtering, we performed a principal component analysis for each population in PLINK. This was to assess if there were discrete subpopulations within the populations and if there were patterns of structure related to ecoregions.

Generation Proxy Selection Mapping (GPSM):

To identify alleles that had changed in frequency over time, we fit a univariate genome-wide linear mixed model (LMM) using GEMMA (Version 0.98.1) [118]. Here, we used the model:

EQUATION 1:

$$y = Xg + Zu + e$$

$$u \sim N(0, G{\sigma_a}^2)$$

$$e \sim N(0, {\sigma_e}^2 I)$$

where **y** is an individual's generation proxy, in our case birth date, and **X** was an incidence matrix that related SNPs to birth dates within each individual and **g** was the estimated effect size for each SNP. An animal's age as of April 5, 2017 was used as the generation proxy in GPSM. We control for confounding population structure, relatedness, and inbreeding with a polygenic term **u** that uses a standardized genomic relationship matrix (GRM) **G** [3] and we estimated ${\sigma_a}^2$ and ${\sigma_e}^2$ using restricted maximum likelihood estimation. Here, continuous age served as a proxy for generation number from the beginning of the pedigree. Other than the tested SNP effects, no fixed effects other than the overall mean were included in the model. We tested each SNP for an association with continuous age. We converted p-values to FDR corrected q-values and used a significance threshold of $q < 0.1$. We performed additional negative-control analyses in each dataset by permuting the date of birth associated with each animal’s genotypes to ensure that the detected GPSM signals were likely to be true positives. Permutation was performed ten times for each population. To visualize the allele frequency history of loci undergoing the strongest selection, we fit a loess and simple linear regressions for date of birth and allele frequencies scored as 0, 0.5 or 1.0 within each individual using R [119]. Results were visualized using ggplot2 [120].

Birth date variance component analysis:

To estimate the amount of variation in birth date explained by GPSM significant SNPs, we performed multi-GRM GREML analyses for birth date in GCTA (v1.92.4) [121]. We built separate GRMs using genome-wide significant markers and all remaining makers outside of significant GPSM loci (> 1 Mb from significant GPSM SNPs to control for markers physically linked to significant GPSM SNPs). To further partition the variance in birth date explained by subsets of SNPs, we performed a GREML analysis using three GRMs created with genome-wide significant ($p < 1\times10^{-5}$) SNPs, an equal number of the next most significant SNPs, and an equal number of unassociated ($p > 0.5$) markers with minor allele frequencies > 0.15, to match the allele frequencies of significant SNPs. These three GRM were each constructed using 268, 548, and 763 SNPs for Red Angus, Simmental, and Gelbvieh, respectively.

Environmental Data:

Thirty-year normals (1981-2010) for mean temperature ((average daily high (°C) + average daily low (°C)/2), precipitation (mm/year), and elevation (m above sea level) for each 4 km^2 of the continental US were extracted from the PRISM Climate Dataset [122], and used as continuous dependent variables in envGWAS analysis. Optimal *K*-means clustering of these three variables grouped each 4 km^2 of the continental US into 9 distinct ecoregions. Using the reported breeder zip code for each individual, we linked continuous environmental variables to animals and partitioned them into discrete environmental cohorts for downstream analysis. For ecoregion assignments, latitude and longitude were rounded to the nearest 0.1 degrees. As a result, some zip codes were assigned to multiple ecoregions. Animals from these zip codes were excluded from the discrete region envGWAS but remained in analyses that used continuous measures as dependent variables.

Environmental Genome-wide Association Studies (envGWAS):

To identify loci segregating at different frequencies within discrete ecoregions or along continuous climate gradients, we used longitudinal environmental data for the zip codes attached to our study individuals as dependent variables in univariate and multivariate genome-wide LMMs implemented in GEMMA (Version 0.98.1). We fit three univariate envGWAS models that used 30-year normal temperature, precipitation, and elevation data as dependent variables. These used an identical model to **EQUATION 1**, but used environmental values as the dependent variable (**y**) instead of birth date. We also fit a combined multivariate model using all three environmental variables to increase power. To identify loci associated with entire climates as opposed to only continuous variables, we fit univariate and multivariate case-control envGWAS analyses using an individual's region assignment described in the "Environmental Data" section as binary phenotypes. Proportion of variation explained (PVE), phenotypic correlations, and genetic correlations were estimated for continuous environmental variables and discrete environmental regions using GEMMA’s implementation of REML.

To ensure that envGWAS signals were not driven by spurious associations, we performed two separate permutation analyses. In the first, we randomly permuted the environmental variables and regions associated with an individual prior to performing each envGWAS analysis, detaching the relationship between an individual’s genotype and their environment. In the second, to ensure that envGWAS signals were not driven by the over-sampling of individuals at particular zip codes, we permuted the environmental variables associated with each zip code prior to envGWAS analysis. These two types of permutation analyses were performed for each dataset and for each type of univariate and multivariate envGWAS analysis. We determined significance using a permutation-derived p-value cutoff ($p < 1 \times 10^{-5}$) [123].

GPSM meta-analyses:

To identify variants undergoing ecoregion-specific allele frequency changes, we performed GPSM analyses within each region with more than 600 individuals. The SNP significance testing effects and standard errors from each of the within-region GPSM analyses were combined into a single meta-analysis for each population using METASOFT (v2.0.1) [124]. We identified loci with high heterogeneity in allele effect size, suggesting region-specific selection. An m-value indicating the posterior-probability of a locus having an effect in a particular ecoregion was calculated for each of these loci [125].

Gene set and tissue set enrichment analysis:

Using the NCBI annotations for the ARS-UCD1.2 *Bos taurus* reference assembly, we located proximal candidate genes near significant SNPs from each of our analyses. We generated two candidate gene lists each from significant GPSM and envGWAS SNPs. Lists contained all annotated genes within 10 kb or 100 kb from significant SNPs. We consolidated significant SNPs from all envGWAS analyses to generate a single candidate gene list for each breed. Using these candidate gene lists, we performed gene ontology (GO) and KEGG pathway enrichment analysis using Clue GO (v2.5.5) [126] implemented in Cytoscape (v3.7.2) [127]. We identified pathways and GO terms where at least two members of our candidate gene list comprised at least 1.5% of the term's total genes. We applied a Benjamini-Hochberg multiple-testing correction to reported p-values and GO terms with FDR corrected p-values < 0.1 were considered significant.

Using the above gene sets, we performed three separate Tissue Set Enrichment Analyses (TSEA) using existing databases of human, mouse, and worm gene expression data. We searched for enriched gene expression with data from the Human Protein Atlas [128] and Mouse ENCODE [129] using the Tissue Enrich tool (v1.0.7) [130]. Additionally, we performed another Tissue Set Enrichment Analysis using GTEx data [131] and a targeted Brain Tissue Set Enrichment Analysis in the pSI R package (v1.1) [132]. Finally, we used Ortholist2 [133] to identify *C. elegans* genes orthologous with members of our envGWAS and GPSM gene lists. We then queried these lists in WormBase's Tissue Enrichment Analysis tool [134,135] to identify specific tissues and neurons with enriched expression in *C. elegans*. We used each tool's respective multiple-testing correction to determine significance. We deemed an enrichment in a tissue “suggestive” when its p-value was < 0.1.

Acknowledgments

We appreciate comments from Wes Warren, Jeremy Taylor, and William Lamberson while writing this manuscript. We also appreciate farmers and ranchers for collecting this data and for data being shared from the Red Angus Association of America, American Simmental Association, and American Gelbvieh Association.

FIGURES

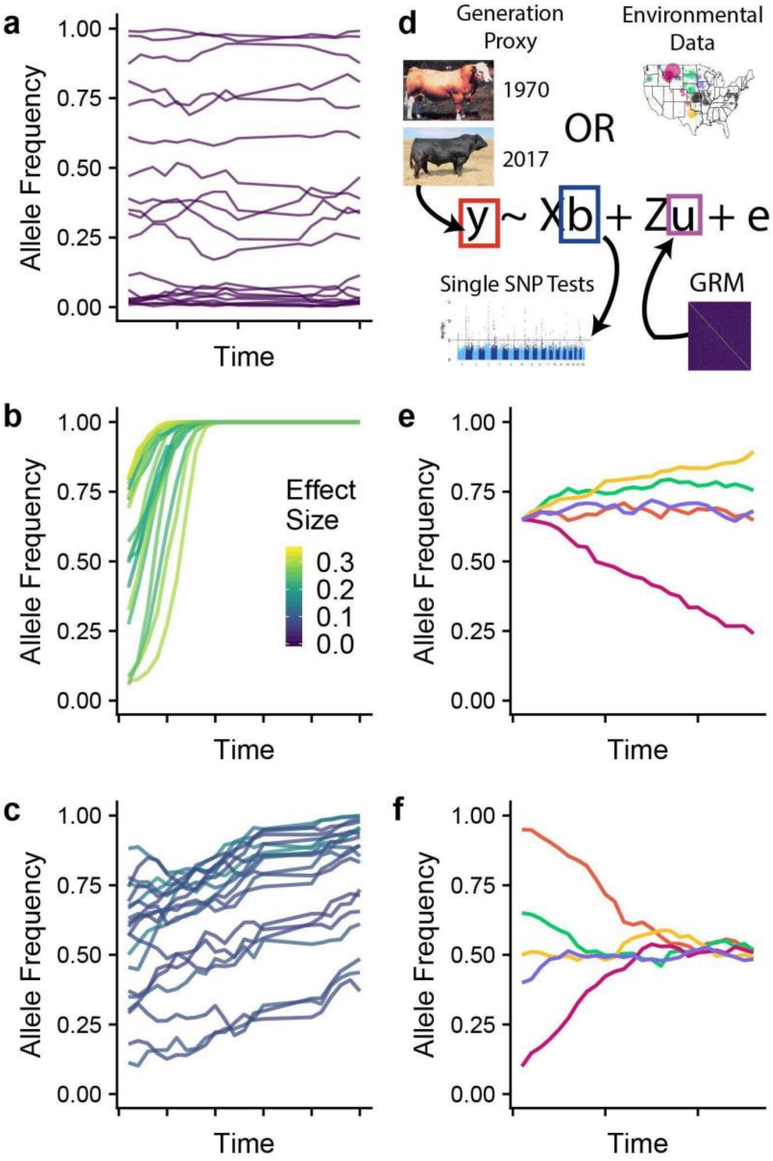


Figure 3.1. Simulated allele frequency trajectories and model overview. (a-c) Allele frequency trajectories for 20 SNPs colored by relative effect sizes from stochastic selection simulations. (a) Effect size = 0, representing stochastic changes in allele frequency due to genetic drift. (b) Large-effect alleles rapidly becoming fixed in the population representing selective sweeps. (c) Moderate-to-small effect size SNPs changing in frequency slowly over time, representing polygenic selection. (d) An

overview of the linear mixed model approach used for Generation Proxy Selection Mapping and environmental GWAS. (e-f) A single SNP under ecoregion-specific selection. Different colors represent the trajectory of a given SNP in one of five different ecoregions. Ecoregion-specific selection can lead to allele frequencies that (e) diverge from or (f) converge to the population mean.

[Figure 1 in text]

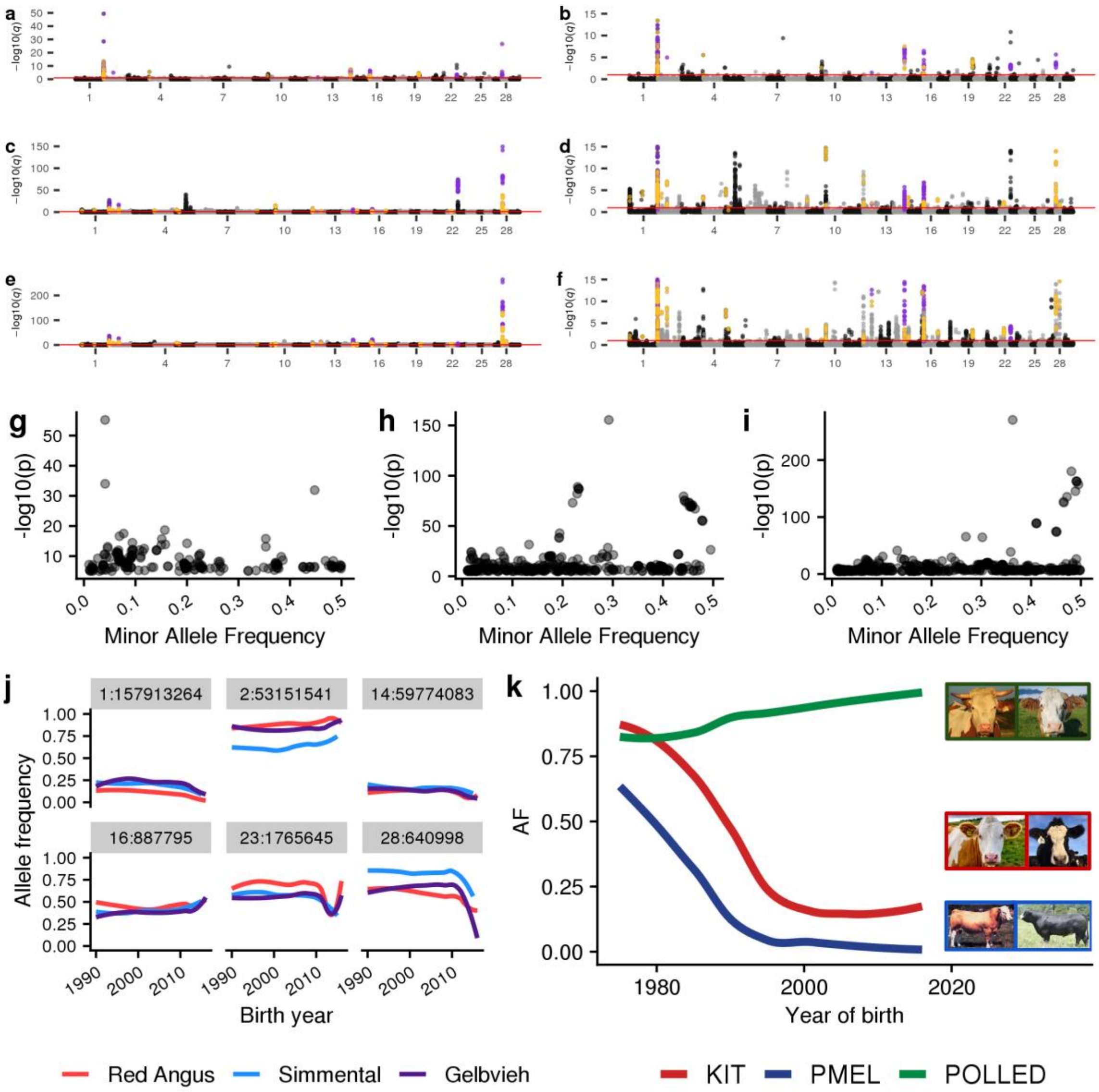


Figure 3.2. Generation Proxy Selection Mapping identifies signals of polygenic selection in three major U.S. cattle populations. Full and truncated (-log$_{10}$(q)< 15) Manhattan plots for GPSM analysis of Red Angus (a & b), Simmental (c & d), and Gelbvieh (e & f). Purple points indicate SNPs significant in all three population-specific GPSM analyses and orange points indicate SNPs significant in two. Minor allele frequency plotted versus -log10(p) values for significant SNPs in (g) Red Angus, (h)

Simmental, and (i) Gelbvieh populations. (j) Smoothed allele frequency histories for the six most significant loci identified as being under selection in all three datasets. (k) Allele frequency histories for three known Mendelian loci that control differences in visual appearance between introduced European and modern US Simmental cattle.

[Figure 2 in text]

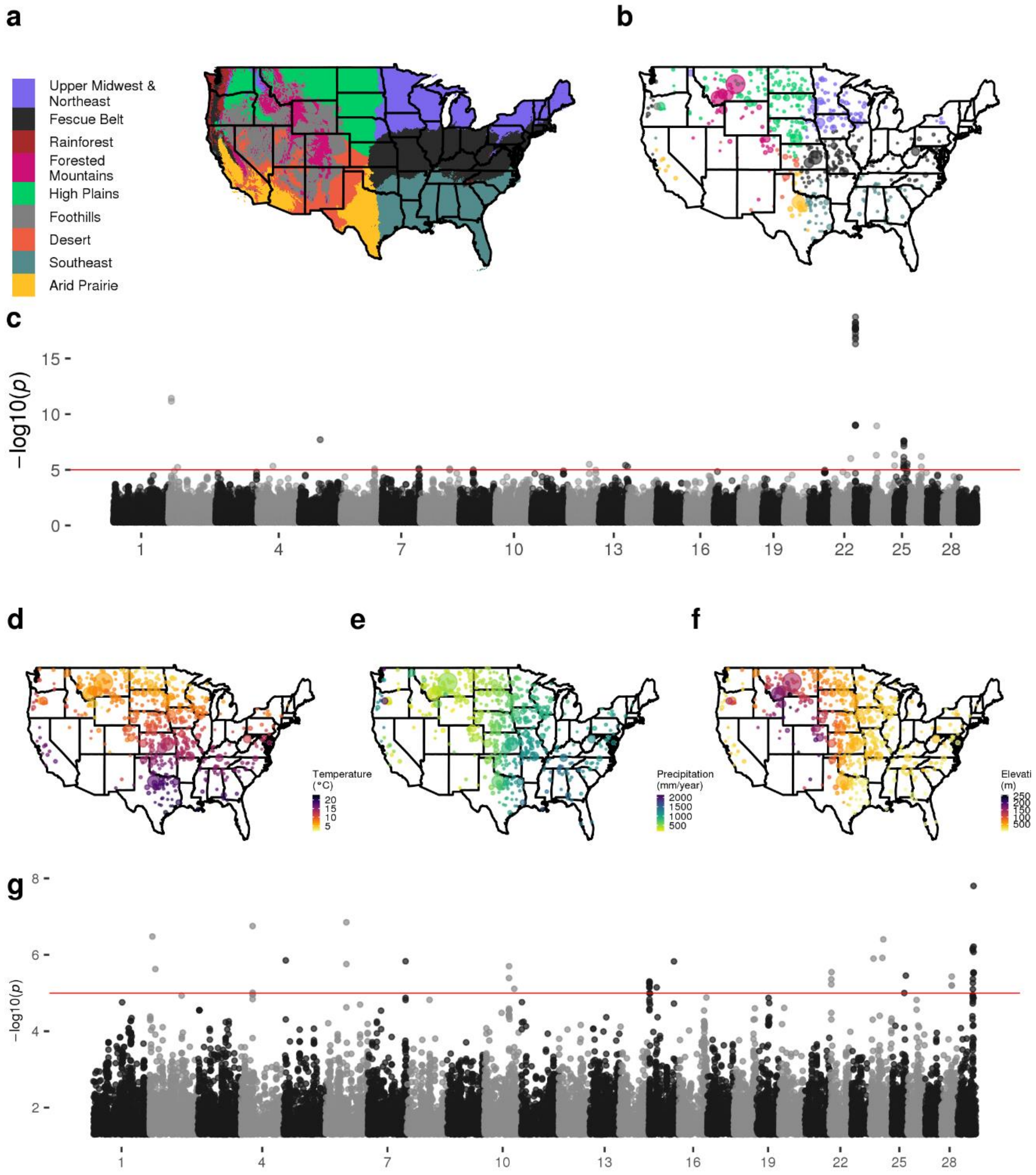


Figure 3.3. Manhattan plots for discrete and continuous envGWAS in Red Angus cattle. (a) Nine continental US ecoregions defined by *K*-means clustering of 30-year normal temperatures, precipitations, and elevations. (b) Locations of sampled Red Angus animals coloured by breeder's ecoregion and sized by the number of animals at that location. (c) Multivariate discrete envGWAS (case-control for six regions with > 600

animals). Locations of sampled Red Angus animals colored by (d) 30-year normal temperature, (e) 30-year normal precipitation, and (f) elevation. (g) Multivariate continuous envGWAS with temperature, precipitation, and elevation as dependent variables. For all Manhattan plots the red line indicates the empirically-derived p-value significance threshold from permutation analysis ($p < 1 \times 10^{-5}$).

[Figure 3 in text]

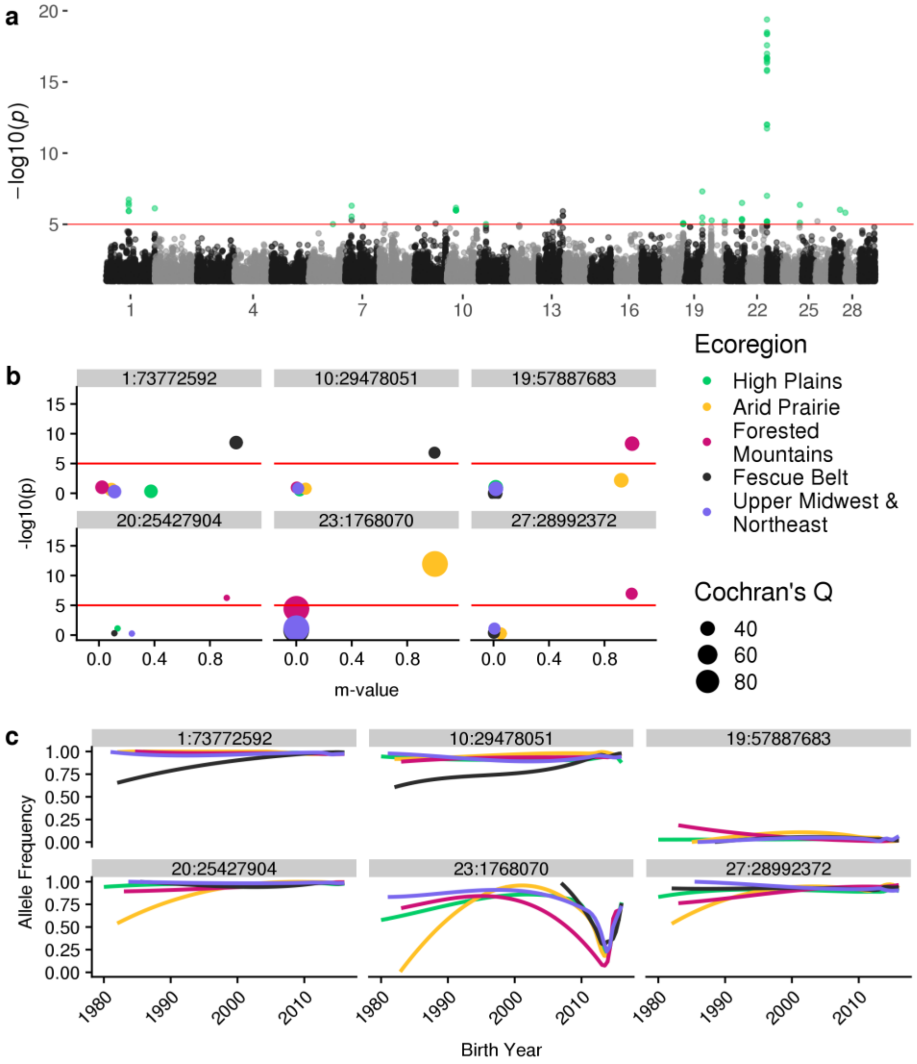


Figure 3.4. Meta-analysis of within-ecoregion GPSM for Red Angus cattle. (a) Manhattan plot of per-variant Cochran's Q p-values. Points coloured green had significant Cochran's Q ($p < 1\times10^{-5}$) and were significant in at least one within-region GPSM analysis ($p < 1\times10^{-5}$). (b) Ecoregion effect plots for lead SNPs from six loci from (a). Points are coloured by ecoregion and are sized based on Cochran's Q value. (c) Ecoregion-specific allele frequency histories for SNPs from (b), coloured by ecoregion. [Figure 4 in text]

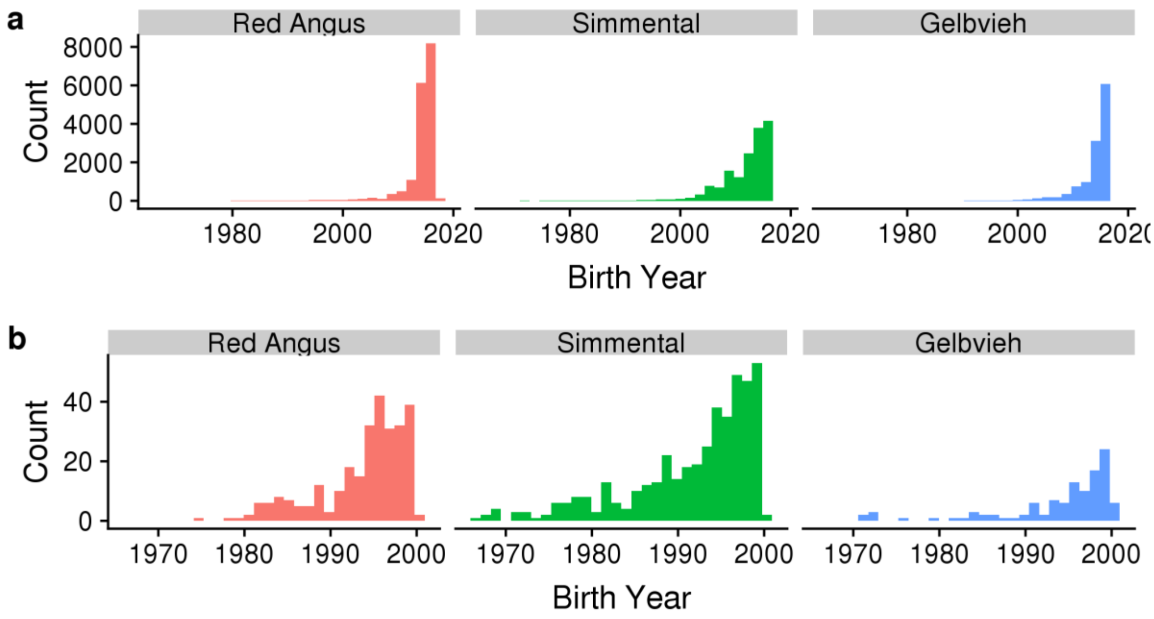


Figure 3.5. Distributions of continuous birth date in sampled Red Angus, Simmental, and Gelbvieh populations. (a) Birth date histograms for complete datasets. (b) Histograms of animal birth dates born before 2000.

[Figure S1 in text]

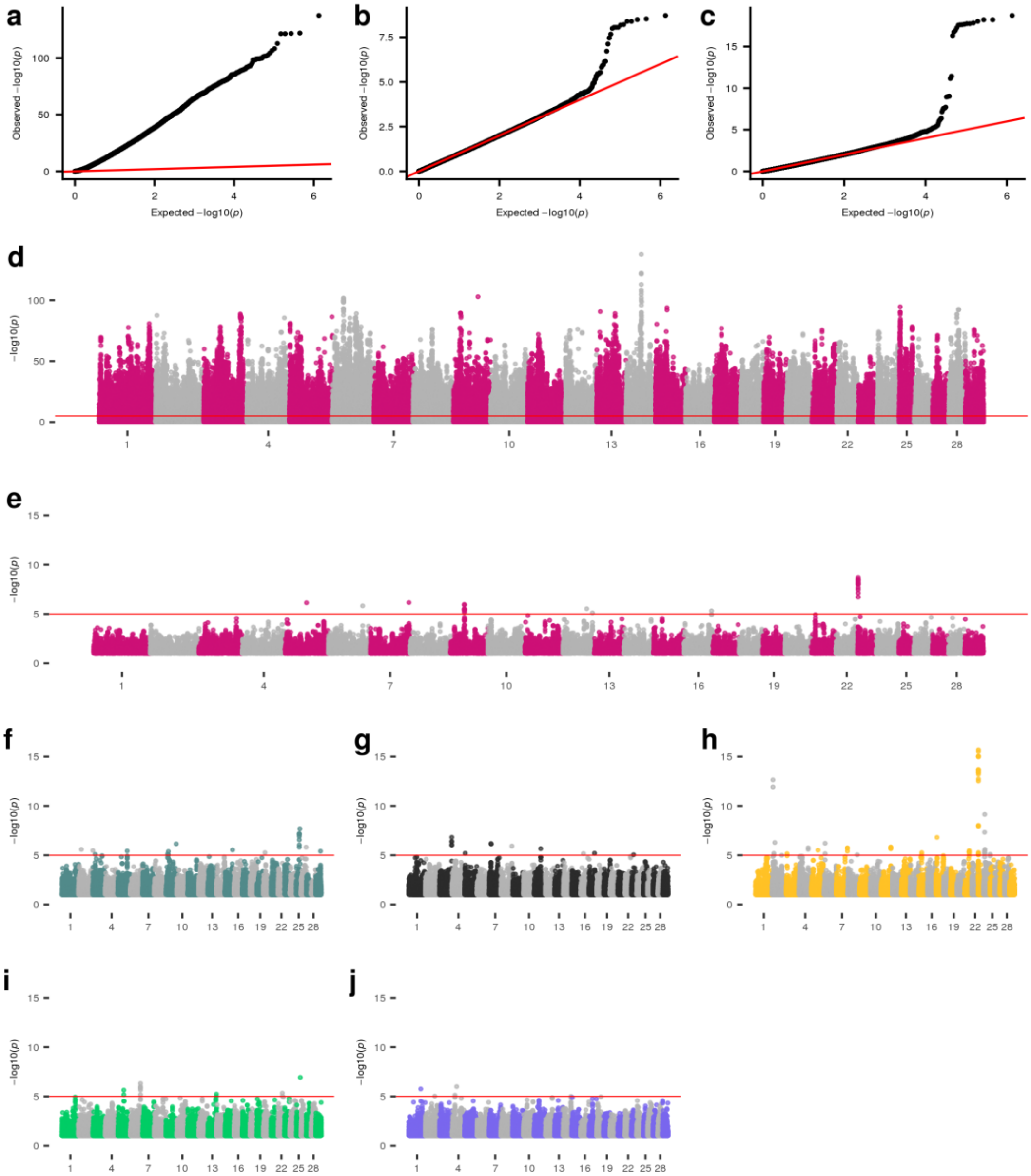


Figure 3.6. Manhattan plots of discrete envGWAS in Red Angus cattle. Q-Q plots for envGWAS p-values of (a) a linear model for Forested Mountains ecoregion membership, (b) a linear mixed model for Forested Mountain ecoregion membership, and (c) a multivariate linear mixed model of ecoregion membership. Univariate discrete

envGWAS for (d) Forested Mountain linear model, (e) Forested Mountains linear mixed model, (f) Southeast, (g) Fescue Belt, (h) Arid Prairie, (i) High Plains, and (j) Upper Midwest & Northeast ecoregions. In all Manhattan plots the red line indicates an empirically-derived p-value significance threshold from permutation testing ($p < 1×10-5$). Note the drastically inflated p-values from the linear model in (a). Further, note that associated loci are not consistent between linear model and linear mixed model, highlighting the need to control for geographic dependency with a genomic relationship matrix.

[Figure S2 in text]

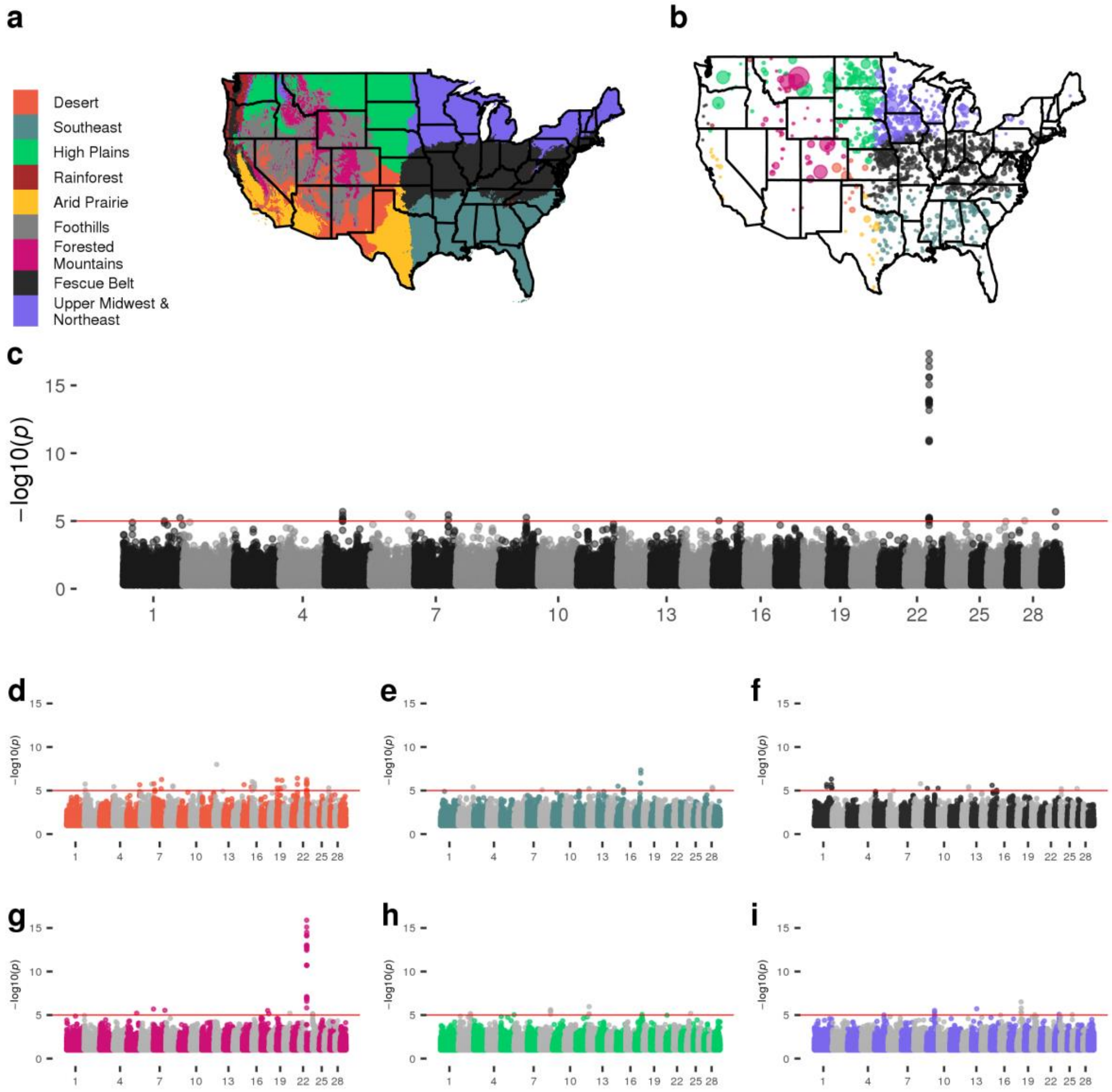


Figure 3.7. Manhattan plots of discrete envGWAS in Simmental cattle. (a) Nine ecoregions of the continental United States defined by *K*-means clustering of 30-year normal temperature, precipitation, and elevation. (b) Locations of Simmental animals colored by breeder’s ecoregion and sized by number of animals at that location. (c) Multivariate envGWAS (case-control for regions with > 600 animals). Univariate discrete envGWAS for (d) Desert, (e) Southeast, (f) Fescue Belt, (g) Forested Mountains,

(h) High Plains, and (i) Upper Midwest & Northeast ecoregions. In all Manhattan plots the red line indicates an empirically-derived p-value significance threshold from permutation testing ($p < 1\times10^{-5}$).

[Figure S3 in Text]

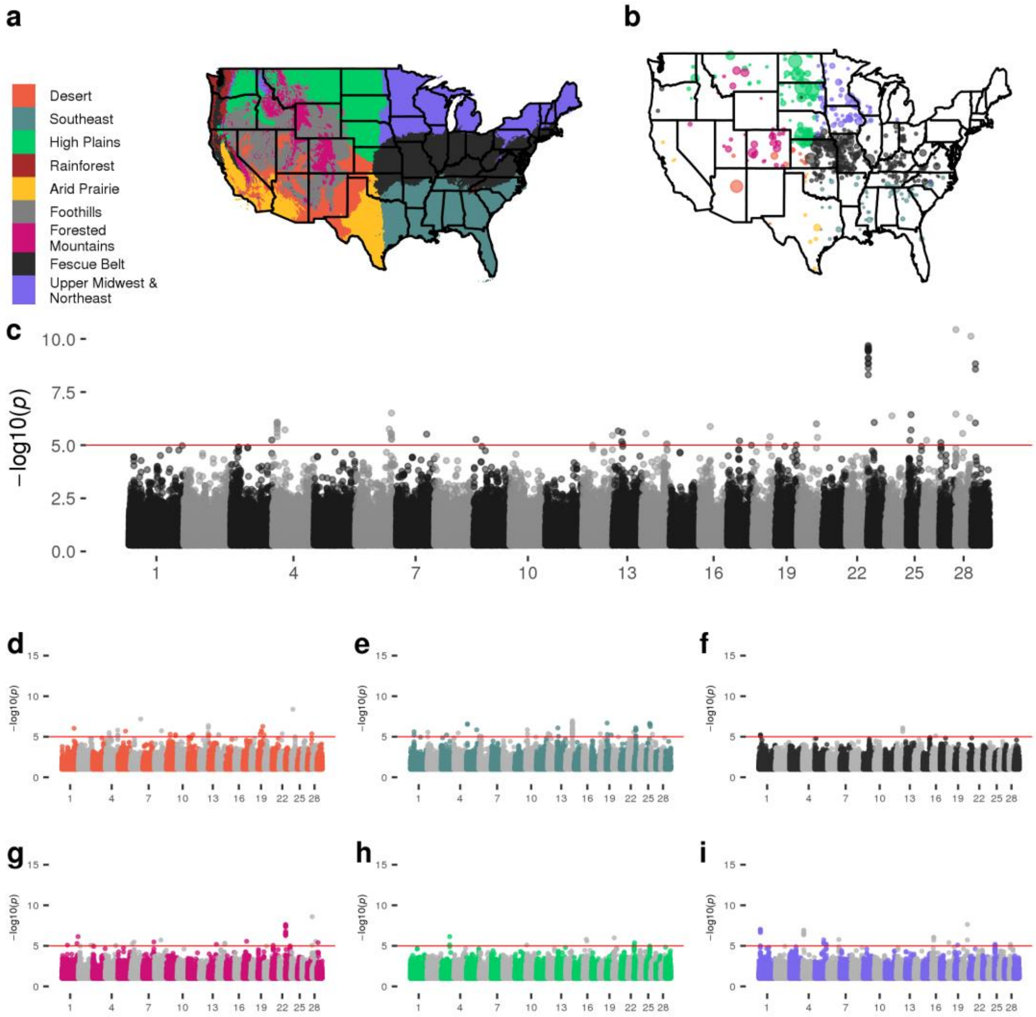


Figure 3.8. Manhattan plots of discrete envGWAS in Gelbvieh cattle. (a) Nine ecoregions of the continental United States defined by *K*-means clustering of 30-year normal temperature, precipitation, and elevation. (b) Locations of Gelbvieh animals colored by breeder's ecoregion and sized by number of animals at that location. (c) Multivariate envGWAS (case-control for regions with > 600 animals). Univariate discrete envGWAS for (d) Desert, (e) Southeast, (f) Fescue Belt, (g) Forested Mountains, (h) High Plains, and (i) Upper Midwest & Northeast ecoregions. In all Manhattan plots the red line indicates an empirically-derived p-value significance threshold from permutation testing ($p < 1\times10^{-5}$). [Figure S4 in text]

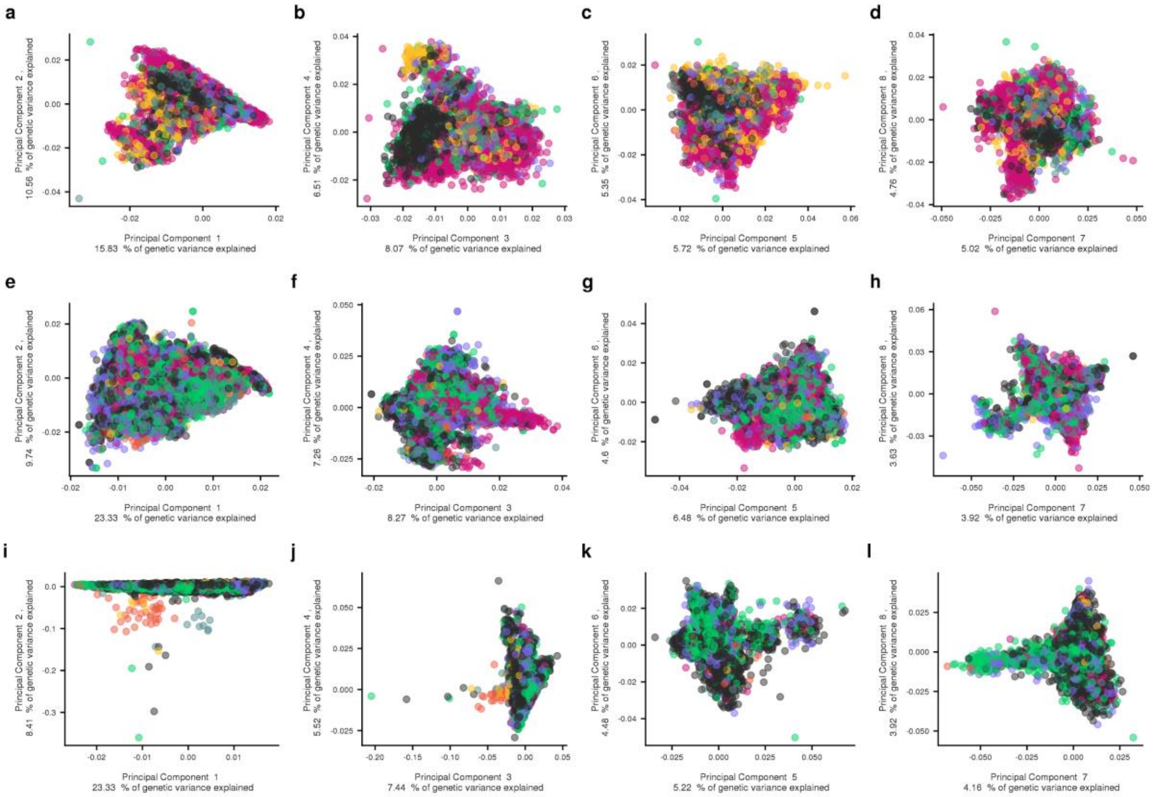


Figure 3.9. Plots of first eight principal components from PCA analysis. Plots for Red Angus (A-D), Simmental (E-H), and Gelbvieh (I-L). Points indicate individuals, colored by their assigned ecoregion.

[Figure S5 in text]

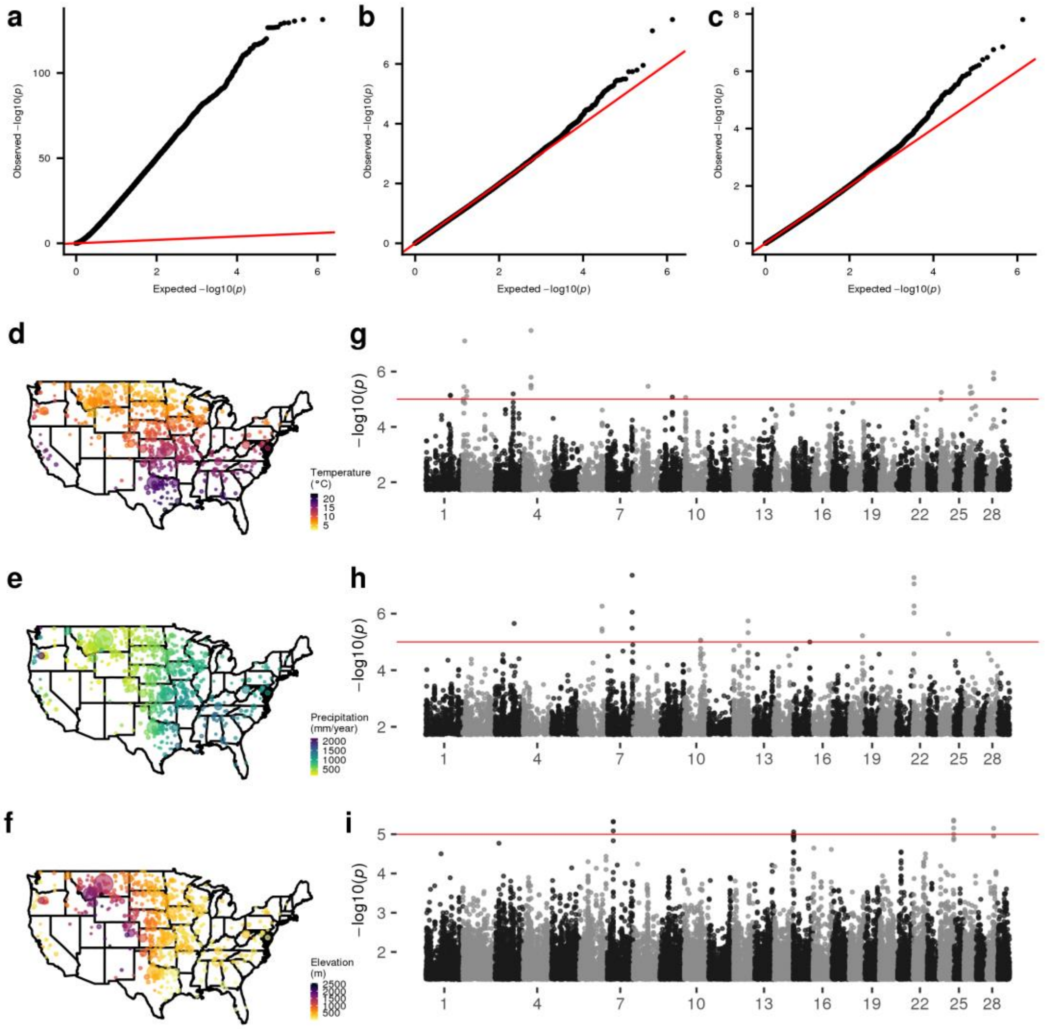


Figure 3.10. Continuous environmental variable envGWAS in Red Angus cattle. Q-Q plots for envGWAS p-values of (a) a linear model for temperature, (b) a linear mixed model for temperature, and (c) a multivariate linear mixed model of temperature, precipitation, and elevation. Geographic distributions colored by (d) temperature, (e) precipitation, (f) elevation. Manhattan plots for univariate envGWAS analysis of (g) temperature, (h) precipitation, (i) elevation. Red lines indicate permutation-derived p-value cutoff of 1×10^{-5}.

[Figure S6 in text]

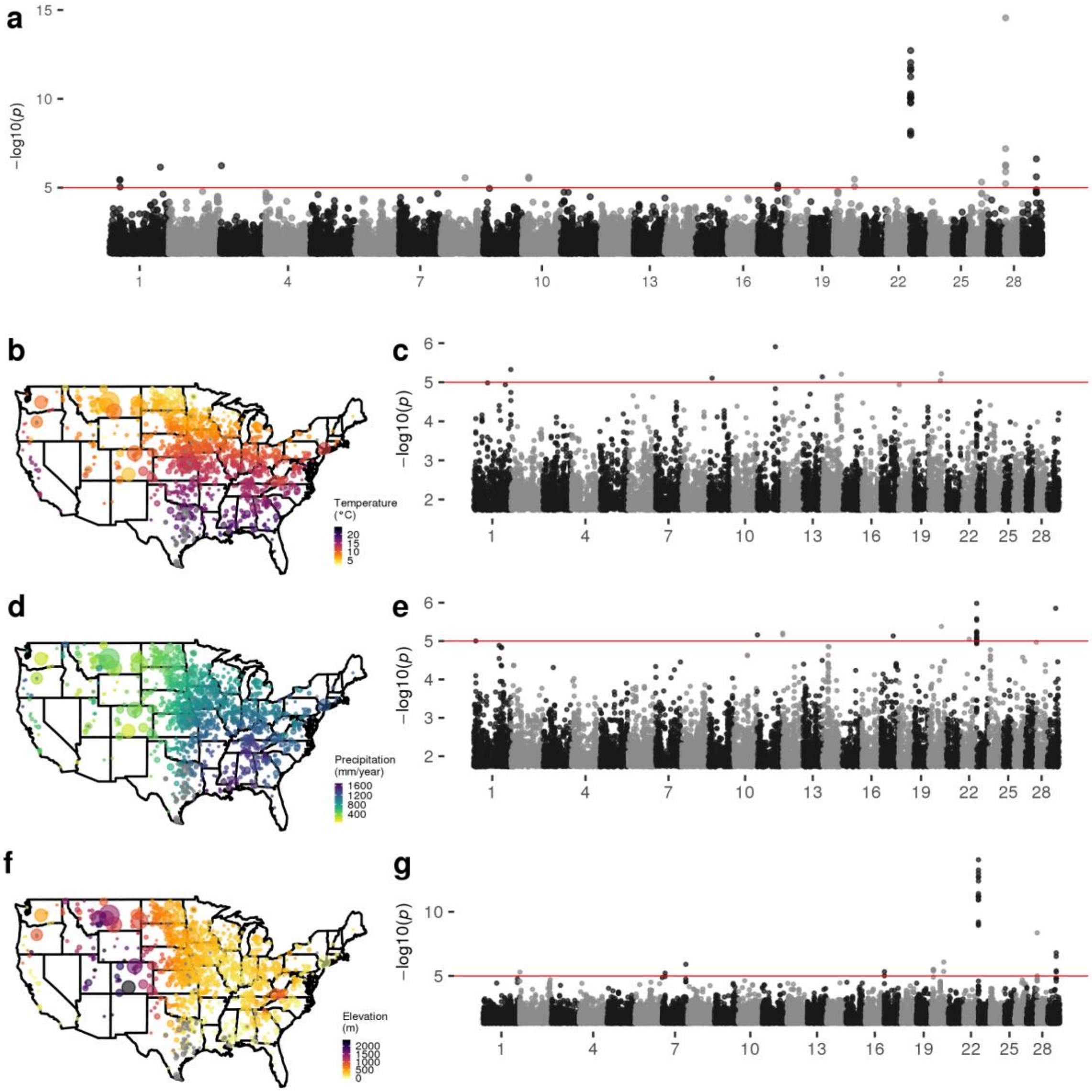


Figure 3.11. Continuous environmental variable envGWAS in Simmental cattle. (a) Multivariate envGWAS of temperature, precipitation, and elevation for Simmental cattle. Geographic distributions colored by (b) temperature, (d) precipitation, (f) elevation. Manhattan plots for univariate envGWAS analysis of (c) temperature, (e) precipitation, (g) elevation. Red lines indicate permutation-derived p-value cutoff of 1×10^{-5}.

[Figure S7 in text]

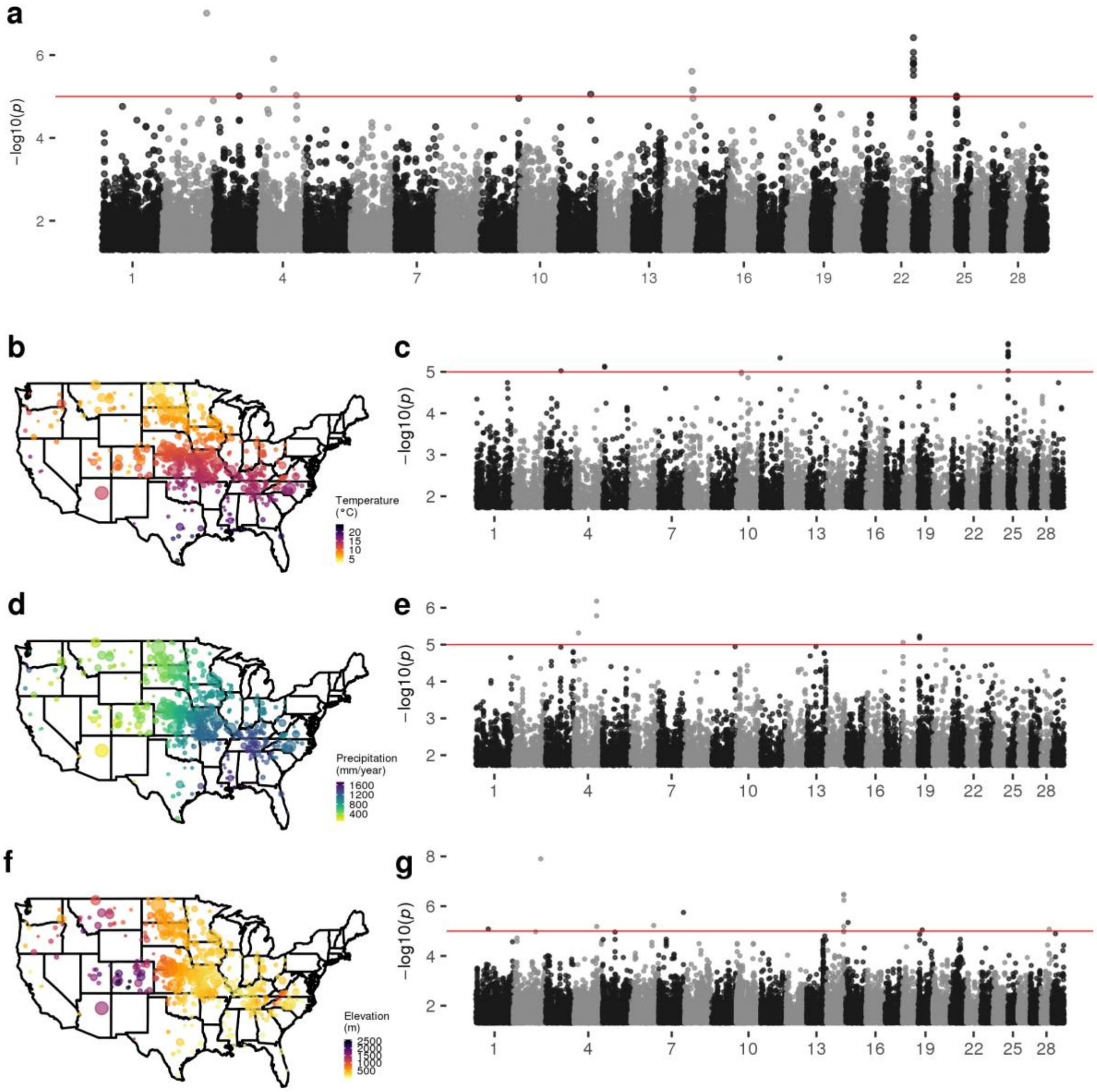


Figure 3.12. Continuous environmental variable envGWAS in Gelbvieh cattle. (a) Multivariate envGWAS of temperature, precipitation, and elevation for Gelbvieh cattle. Geographic distributions colored by (b) temperature, (d) precipitation, (f) elevation. Manhattan plots for univariate envGWAS analysis of (c) temperature, (e) precipitation, (g) elevation. Red lines indicate permutation-derived p-value cutoff of $1×10^{-5}$.

[Figure S8 in text]

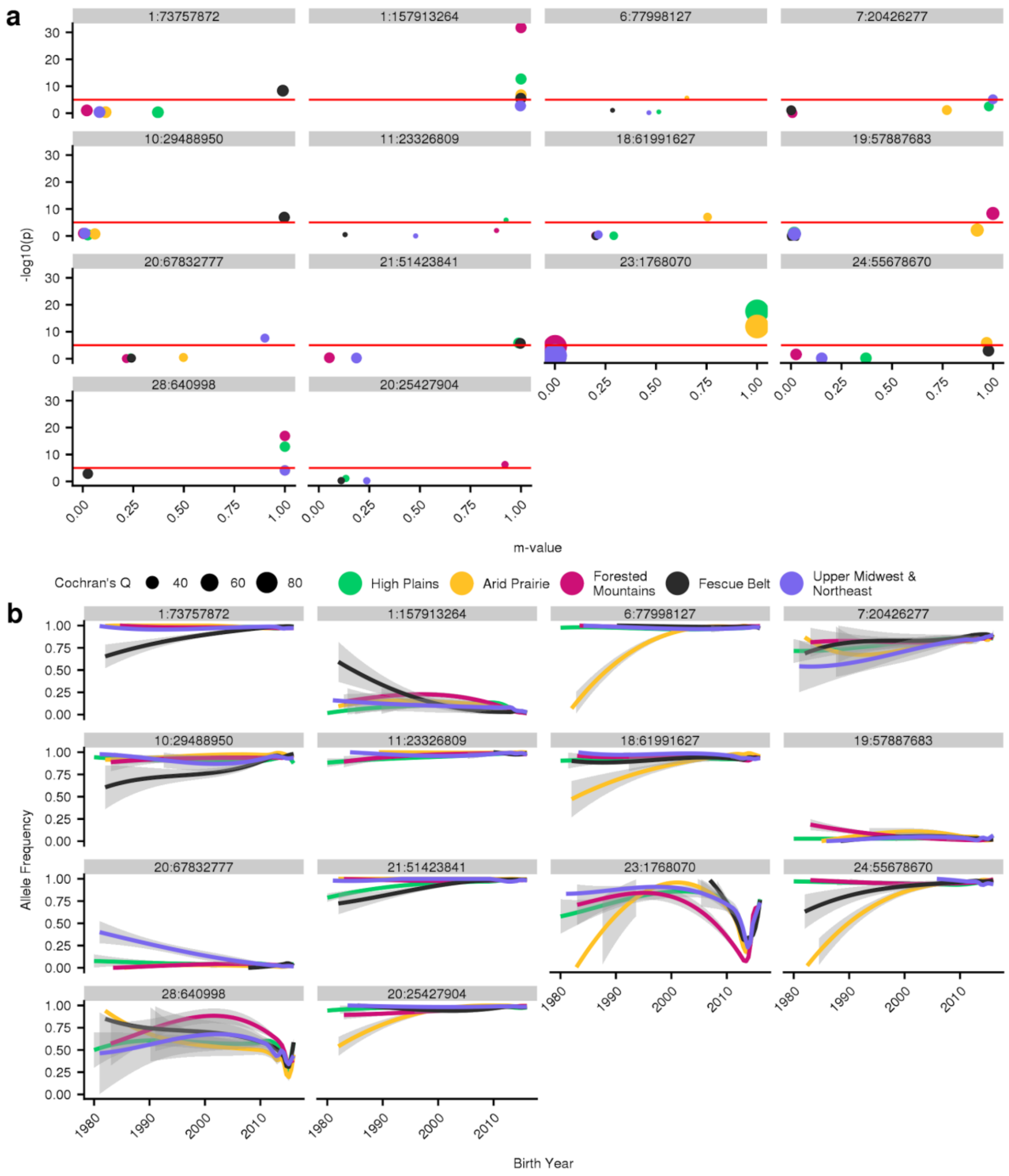


Figure 3.13. PM-plots and region-specific allele frequency trajectories for meta-analysis SNPs of interest in the Red Angus population ecoregions with > 1,000 genotyped animals. (a) PM-plots for lead SNPs of significant within-region GPSM meta-analysis (Cochran's Q p-value > 1×10^{-5} and significant in at least one region-specific GPSM analysis p < 1×10^{-5}). Each box represents the lead SNP, colored by

ecoregion, and sized by Cochran’s Q value (for heterogeneity). (b) Region-specific allele frequency trajectories for lead SNPs since 1980, generated by fitting smoothed loess regression of allele frequency on birth date. Trajectories are colored by ecoregion.

[Figure S9 in text]

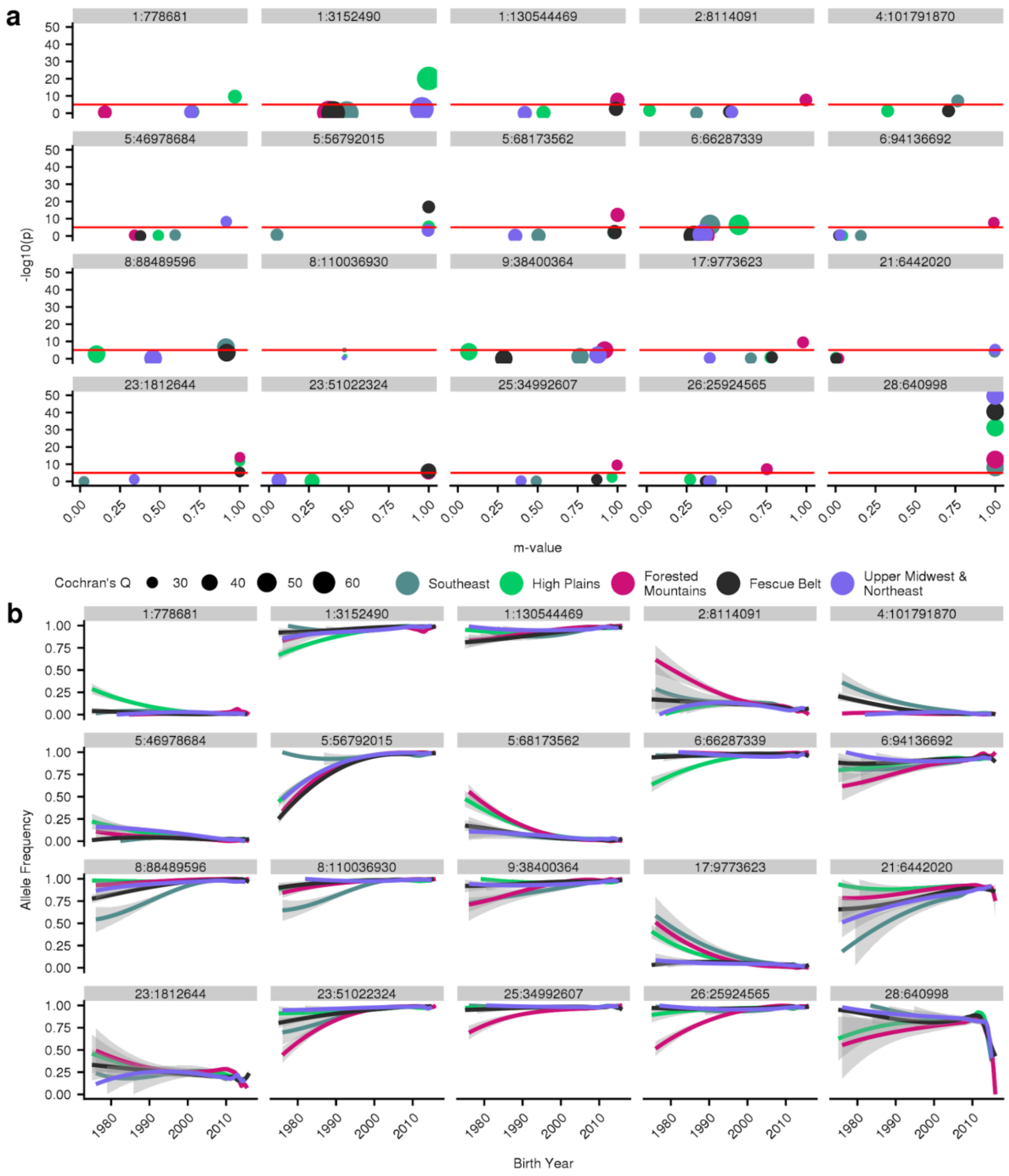


Figure 3.14. PM-plots and region-specific allele frequency trajectories for meta-analysis SNPs of interest in the Simmental population ecoregions with > 1,000 genotyped animals. (a) PM-plots for lead SNPs of significant within-region GPSM meta-analysis (Cochran's Q p-value > 1×10^{-5} and significant in at least one region-

specific GPSM analysis $p < 1\times10^{-5}$). Each box represents the lead SNP, colored by ecoregion, and sized by Cochran's Q value (for heterogeneity). (b) Region-specific allele frequency trajectories for lead SNPs since 1980, generated by fitting smoothed loess regression of allele frequency on birth date. Trajectories are colored by ecoregion.

[Figure S10 in text]

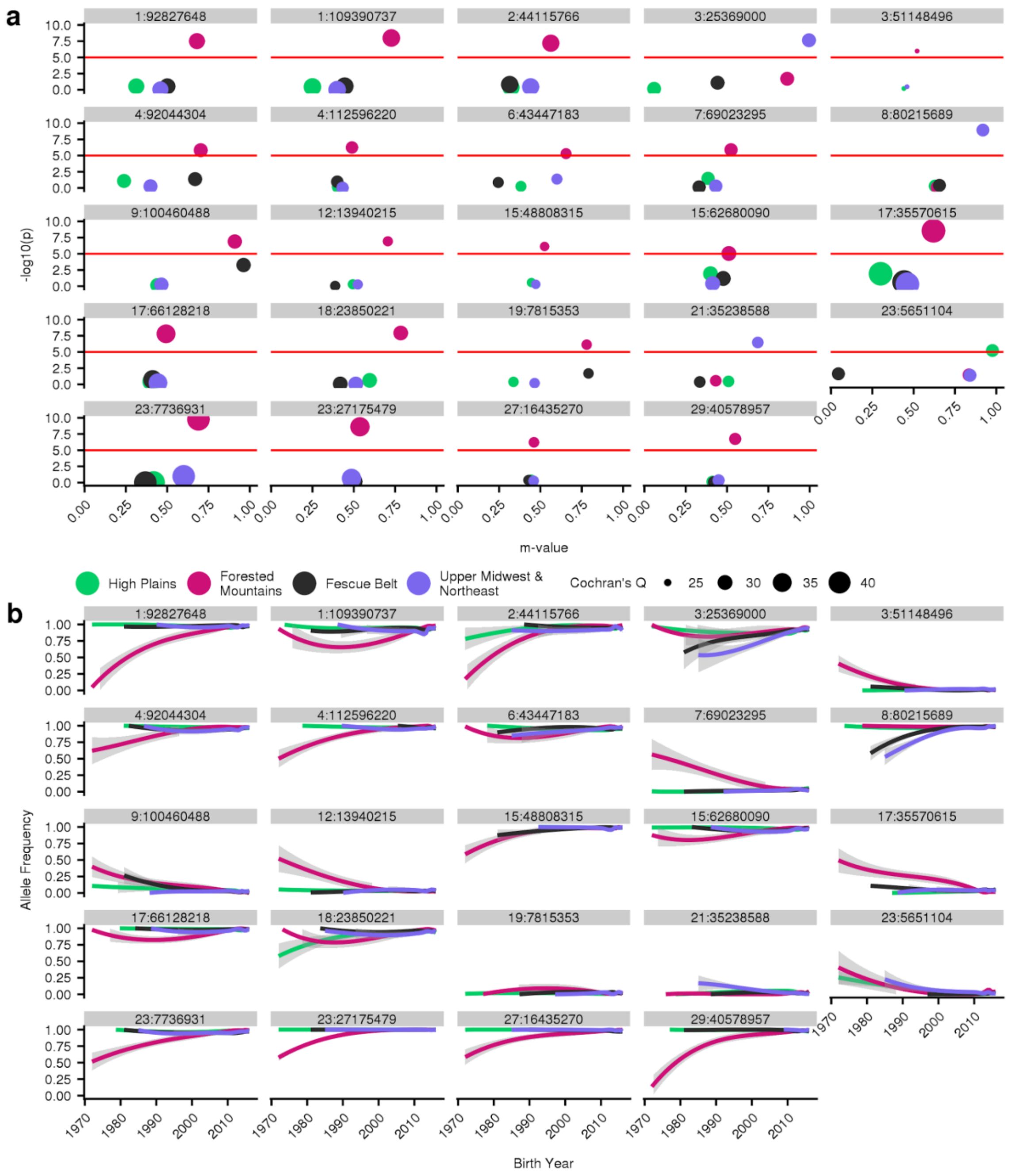


Figure 3.14. PM-plots and region-specific allele frequency trajectories for meta-analysis SNPs of interest in the Gelbvieh population ecoregions with > 1,000 genotyped animals. (a) PM-plots for lead SNPs of significant within-region GPSM meta-analysis (Cochran's Q p-value > 1×10^{-5} and significant in at least one region-specific GPSM analysis p < 1×10^{-5}). Each box represents the lead SNP, colored by

ecoregion, and sized by Cochran’s Q value (for heterogeneity). (b) Region-specific allele frequency trajectories for lead SNPs since 1980, generated by fitting smoothed loess regression of birth of allele frequency on birth date. Trajectories are colored by ecoregion.

[Figure S11 in text]

TABLES

Tables 3.3, 3.6, 3.9, 3.11, 3.16, 3.18, 3.20, 3.21, 3.22 are too large for print, but can be found as tabs in the following spreadsheet:

https://drive.google.com/file/d/1ZMVEYhND9WmvvrBhbyK7IKI3-hU-Im8n/view?usp=sharing

Table 3.1. Variation in birth date explained by three classes of SNPs. The PVE estimates (standard error in parentheses) from a genomic restricted maximum likelihood (GREML) variance component analysis of birth date using three GRMs created from: 1) genome-wide significant SNPs ($q < 0.1$), 2) an equivalent number of the next most significant SNPs outside of genome-wide significant associated regions, and 3) an equivalent number of non-significant SNPs ($p>0.5$) randomly sampled from genomic regions that did not harbor genome-wide significant associations.

Population	**Genome-wide significant SNPs***	**Suggestive significant SNPs***	**Other SNPs***	**Total**
Red Angus	0.187 (0.026)	0.148 (0.015)	0.031 (0.004)	0.366 (0.023)
Simmental	0.239 (0.020)	0.194 (0.014)	0.037 (0.004)	0.470 (0.016)
Gelbvieh	0.225 (0.017)	0.193 (0.013)	0.008 (0.003)	0.426 (0.018)

*Contained 268, 548, and 763 SNPs for Red Angus, Simmental, and Gelbvieh, respectively

[Table 1 in text]

Table 3.2. GPSM and envGWAS gene dropping simulation results. Ten repetitions of a 200,000 SNP gene dropping experiment through the complete Red Angus pedigree. Real GPSM and envGWAS phenotypes were used to identify significant SNPs (p-value < 1×10^{-5}). Multiple nearby SNPs were grouped as “genomic regions" when they were within 1Mb of one another.

Analysis	**Analyzed Individuals**	**Median Number Significant SNPs (SD)**	**Median Significant p-value (SD)**
GPSM	15,315	5 (3.931)	$4.35 \mathrm{x} 10^{-6}$ ($3.08 \mathrm{x} 10^{-6}$)
envGWAS	15,315	5 (5.81)	$3.89 \mathrm{x} 10^{-6}$ ($3.108 \mathrm{x} 10^{-6}$)

[Table S1 in text]

Table 3.3. GPSM stochastic simulation results. Descriptions of 36 selection scenarios, and the corresponding true and false positive rates for GPSM detecting simulated QTL under selection (simulated QTL GPSM p-value < 1×10^{-5}). Each scenario's true and false positive statistics were calculated based on 10 replicates starting with different founder populations and selected randomly (false positives) or based on true breeding value (true positives). For each scenario, we report the number simulated QTL, the number of total crosses performed using 50 males and 500 females, the distribution from which QTL effects were drawn from, and the number of generations of selection performed. In each case, 10,000 simulated individuals were randomly chosen to be genotyped (evenly each generation).

https://drive.google.com/file/d/1ZMVEYhND9WmvvrBhbyK7IKI3-hU-Im8n/view?usp=sharing

[Table S2 in text] [Tab S2 in linked spreadsheet above]

Table 3.4. GPSM datasets from three major U.S. beef cattle populations. Sample sizes are reported prior to and after filtering on individual call rate individuals with reported birth dates.

Breed	**Sample Size (After Filtering)**	**Median Birth Date**	**Mean Birth Date**	**Min Birth Date**	**Max Birth Date**
Red Angus	16,331 (15,295[1])	2014-10-16	2013-12-26	1975-03-21	2017-04-23
Simmental	17,468 (15,350)	2013-08-09	2011-08-11	1966-02-16	2016-04-06
Gelbvieh	12,563 (12,031)	2015-01-11	2013-09-23	1970-12-02	2016-09-21

[1] After removing non-purebred animals

[Table S3 in text]

Table 3.5. The proportion of variation in birth date explained (PVE) by markers in GPSM analysis. PVE calculated for each population dataset in full, and subsetted to individuals born within the last 20 or 10 years. The standard errors of PVE estimates are reported in parentheses.

Population	**Full Dataset PVE (se)**	**20-year PVE (se)**	**10-year PVE (se)**	**Shuffled PVE (se)**
Red Angus	0.520 (0.013)	0.358 (0.013)	0.406 (0.014)	9.82×10^{-6} (0.002)
Simmental	0.588 (0.009)	0.551 (0.010)	0.401 (0.014)	9.70×10^{-4} (0.003)
Gelbvieh	0.459 (0.015)	0.454 (0.015)	0.361 (0.014)	5.05×10^{-4} (0.004)

[Table S4 in text]

Table 3.6. Summary statistics and candidate genes from significant GPSM SNPs for Red Angus, Simmental, and Gelbvieh populations. Variants are significant if GPSM q-value < 0.1. Genomic locations are reported based on coordinates from ARS-1.2 genome assembly. Candidate genes were assigned to a SNP if within 10 kb of a significant SNP.

https://drive.google.com/file/d/1ZMVEYhND9WmvvrBhbyK7IKI3-hU-Im8n/view?usp=sharing

[Table S5 in text] [Tab S5 in linked spreadsheet above]

Table 3.7. Summary statistics of allele frequency change (ΔAF) per generation for significant GPSM SNPs. ΔAF is the slope of a simple regression of allele frequency on birth date multiplied by a generation interval of 5 years.

Breed	N SNPs (GPSM q < 0.1)	Mean ΔAF per generation (sd)	Median ΔAF per generation	Min ΔAF per generation	Max ΔAF per generation
Red Angus	268	0.018 (0.011)	0.017	4.97×10^{-5}	0.076
Simmental	548	0.024 (0.017)	0.022	6.79×10^{-5}	0.093
Gelbvieh	762	0.033 (0.028)	0.024	1.01×10^{-4}	0.223

[Table S6 in text]

Table 3.8. Significant GPSM variants identified in at least two populations. Lead SNPs from significant loci identified in GPSM analyses of Red Angus, Simmental, and Gelbvieh cattle populations. Locus reported if it was identified in GPSM analysis of at least two populations. Candidate gene is the annotated gene closest to lowest p-value SNP in peak if < 200 kb away. Associations are from cattle literature unless otherwise reported.

CHR	POS	Nearest Candidate Gene(s) (Distance)	Known Candidate Gene Associations	References	Datasets
1	157,913,264	LOC112448253 (within)	lncRNA, PRDM9		ALL
2	53,151,541	*ARHGAP15* (within)	Immune functions, Trypanosomiasis resistance	[136–138]	ALL
12	46,730,506	*DACH1* (182.1 kb)	Feed efficiency/growth	[139]	ALL
14	59,774,083	*LRP12* (261.5 kb), LOC112449532 (113.6 kb)	(*LRP12*) Feed efficiency/growth	[53,140]	ALL
16	955,146	*ADORA1* (within), *MYBPH* (2.0 kb)	Fertility/immune (*ADORA1*)/muscle growth (*MYBPH*)	[141–143]	ALL
23	1,768,070	LOC782044 (25.7 kb)			ALL
28	640,998	*RHOU* (56.3 kb), Olfactory gene cluster (166.5kb)	Bone development, Innate immune system	[144]	ALL
1	3,144,864	URB1 (within)	Embryonic lethal (pigs)	[145]	SIM/GEL
1	73,642,609	> 200 kb to a gene			SIM/GEL

3	119,429,700	CSF2RA (within)	TB Resistance	[146]	RAN/SIM
5	5,266,489	ENSBTAG00000050164 (within)	lncRNA		SIM/GEL
8	113,265,058	> 200 kb to a gene			SIM/GEL
9	78,704,840	PWWP2B (23.6 kb)	Bovine fetus muscle expression	[147]	RAN/GEL
9	104,228,150	PDCD2 (111 kb)			SIM/GEL
10	940,793	MCC (within)			SIM/GEL
12	518,595	PCDH20 (37.7 kb)			SIM/GEL
15	77,106,772	DDB2 (within)	Calving Ease	[148]	RAN/GEL
15	84,680,379	LOC617614	Olfactory Receptor		SIM/GEL
16	78,129,493	> 200 kb to a gene			SIM/GEL
19	53,947,810	BIRC5 (within)	RFI (DE), Fertility and Reproduction	[149]	RAN/GEL
22	24,625,216	Between CNTN6 and CNTN4	Known Milk Yield QTL	[150]	SIM/GEL
28	20,762,645	> 200 kb to a gene			SIM/GEL

[Table S7 in text]

Table 3.9. Gene enrichment analysis of GPSM candidate genes in Red Angus, Simmental, and Gelbvieh populations. Candidate genes were annotated genes < 10 kb to significant GPSM SNPs (q < 0.1). Significant (FDR-corrected p-values < 0.1) KEGG pathways and GO biological processes are reported for each breed.

https://drive.google.com/file/d/1ZMVEYhND9WmvvrBhbyK7IKI3-hU-Im8n/view?usp=sharing

[Table S8 in text] [Tab S8 in linked spreadsheet above]

Table 3.10. TissueEnrich analysis using GPSM gene sets from Red Angus, Simmental, and Gelbvieh populations. TSEA from TissueEnrich software using Human Protein Atlas gene expression data. Enrichment analysis carried out for candidate genes within 10 kb of significant envGWAS SNPs. For each test, we report the number of tissue specific genes, their average fold change, and the FDR-corrected $\log_{10}$ p-value for tissue enriched expression.

https://drive.google.com/file/d/1ZMVEYhND9WmvvrBhbyK7IKI3-hU-Im8n/view?usp=sharing

[Table S9 in text] [Tab S9 in above linked spreadsheet]

Table 3.11. Tissue enrichment analysis results from GPSM gene sets in Red Angus, Simmental, and Gelbvieh populations using the pSI R package and human GTEx expression data. Enrichment significance values for four specificity index thresholds (pSI) of 25 human tissues types. Each combination of stringency for enrichment (pSI) and tissue reports a p-value for Fisher's Exact Test and a Benjamini Hochberg corrected p-value reported in parentheses. Tissue-gene-set combinations that are significant (Benjamini-Hochberg p-value < 0.1) are highlighted in red, those that are suggestive (raw p-value < 0.1) are highlighted in green.

https://drive.google.com/file/d/1ZMVEYhND9WmvvrBhbyK7IKI3-hU-Im8n/view?usp=sharing

[Table S10 in text] [Tab S10 in above linked spreadsheet]

Table 3.12. Ecoregion distribution of Red Angus, Simmental, and Gelbvieh populations. Counts of analyzed individuals in each region for each dataset after filtering and region assignment based on individual's breeder zip code.

Region	**Red Angus**	**Simmental**	**Gelbvieh**
Desert	367[1]	321[1]	408[1]
Southeast	615[1]	1,073[1,2]	422[1]
High Plains	2,796[1,2]	3,645[1,2]	4,022[1,2]
Rainforest	0	62	1
Arid Prairie	1,206[1,2]	181	87
Foothills	136	0	0
Forested Mountains	4,525[1,2]	2,589[1,2]	704[1,2]
Fescue Belt	3,011[1,2]	4,393[1,2]	4,482[1,2]
Upper Midwest & Northeast	1,513[1,2]	2,524[1,2]	1,072[1,2]
Total	**14,169**	**14,788**	**11,198**

[1] included in multivariate discrete envGWAS analysis

[2] included in "large region" multivariate envGWAS analysis

[Table S11 in text]

Table 3.13. Univariate REML estimates of PVE for continuous environmental variables in genotyped Red Angus, Simmental, and Gelbvieh populations. Standard errors for PVE estimates are reported in parentheses.

Variable	**Red Angus PVE (se)**	**Simmental PVE (se)**	**Gelbvieh PVE (se)**
Temperature	0.597 (0.010)	0.586 (0.011)	0.691 (0.010)
Precipitation	0.526 (0.011)	0.602 (0.011)	0.677 (0.010)
Elevation	0.594 (0.010)	0.585 (0.011)	0.644 (0.011)

[Table S12 in text]

Table 3.14. Univariate estimates of PVE for discrete ecoregion assignment in genotyped Red Angus, Simmental, and Gelbvieh populations. Standard errors for PVE estimates are reported in parentheses.

Variable	**Red Angus PVE (se)**	**Simmental PVE (se)**	**Gelbvieh PVE (se)**
Desert	0.646 (0.010)	0.517 (0.012)	0.726 (0.010)
Southeast (SE)	0.408 (0.010)	0.547 (0.013)	0.478 (0.013)
High Plains (HP)	0.641 (0.011)	0.588 (0.010)	0.694 (0.010)
Arid Prairie (AP)	0.463 (0.011)	0.566 (0.014)	NA
Forested Mountains (FM)	0.575 (0.011)	0.594 (0.010)	0.615 (0.013)
Fescue Belt (FB)	0.673 (0.010)	0.545 (0.011)	0.649 (0.011)
Upper Midwest & Northeast (UMWNE)	0.548 (0.012)	0.509 (0.012)	0.609 (0.013)

[Table S13 in text]

Table 3.15. Candidate genes for discrete ecoregion multivariate envGWAS in Red Angus cattle. Lead SNP in envGWAS peak is reported along with nearest plausible candidate genes (provided < 250 kb from lead SNP). If association was also identified in univariate analysis, it is reported. Potentially adaptive associations are reported along with references.

CHR	POS	Nearest Candidate Gene(s) (Distance)	Univariate Continuous Association	Univariate Ecoregion Association	Candidate Gene Adaptive Associations	Reference
2	7,571,508	DIRC1 (110.3 kb), COL5A2 (212.7 kb)	Multivariate, Temperature	FM	Sweep region (human), blood pressure (human)	[17,151,152]
2	25,519,381	GORASP2 (4.47 kb)	N/A	AP	Adaptive signature (fish)	[153]
4	43,928,337	MAGI2 (within)	N/A	UMWNE	Selection signature, imprinted (cattle)	[154,155]
5	59,498,938	LOC788524 (OR9K2) (1.42 kb) [Olfactory cluster]	N/A	FM	Olfactory receptor cluster, selection signature (bison)	[156]
6	96,217,506	RASGEF1B (within)	N/A	HP	Immune function, adaptation signature (human), Response to viral infections.	[157,158]
8	84,146,584	CENPP (within)	N/A	N/A	African cattle CNV, hypoxia (human)	[159,160]
12	58,438,020	N/A	N/A	N/A		

13	75,784,164	ZMYND8 (within)			Immune function, DNA damage repair	[161,162]
13	81,716,869	PFDN4 (71.97 kb)	N/A	HP	Dermatitis (humans)	[163]
22	48,873,180	DUSP7 (40.83 kb)			Heat stress/sperm motility (cattle), Stress response	[164,165]
23	1,768,070			AP, FM		
24	10,628,280	CDH7 (151.7 kb)	MV	AP	Developmental processes	[166]
24	62,228,300	LOC100298064 (HAVCR1) (15.2kb)		AP	Immune function	[167]
25	26,511,450	SPN (CD43) (within)	N/A	SE	MHC-Class I, Immune response (cattle)	[168]
25	35,085,041	CUX1 (67.8 kb)	Multivariate	N/A	Hair phenotype (goats, mice)	[100,101]
26	34,432,722	ADRB1 (81.5 kb)	Temperature		Selection (sporting dogs), climate adaptation (lizards)	[169,170]

[Table S14 in text]

Table 3.16. envGWAS significant SNPs and candidate genes. SNPs were significant when $p < 1\times10^{-5}$. Candidate genes are genes within < 10 kb of significant envGWAS SNPs. We report significant SNPs from all univariate and multivariate analyses for both continuous environmental variables and discrete environments in Red Angus, Simmental, and Gelbvieh populations.

https://drive.google.com/file/d/1ZMVEYhND9WmvvrBhbyK7IKI3-hU-Im8n/view?usp=sharing

[Table S15 in text] [Tab S15 in above linked spreadsheet]

Table 3.17. Candidate genes identified in multivariate envGWAS analyses using continuous environmental attributes as dependent variables. Chromosome and genomic positions are for lead SNP in peak. Closest gene is identified as a candidate (if < 250 kb from lead SNP).

CHR	POS	Nearest Annotated Gene (Distance)	Univariate association (if $p < 1×10^{-5}$)	Adaptation- related associations	References	Breed(s)
2	7,571,508	DIRC1 (110.3 kb), COL5A2 (212.7 kb)	Temperature	Sweep region (human), blood pressure (human)	[17,151]	RAN
2	16,014,918	No genes < 200 kb	Temperature			RAN
4	32,945,494	ABCB1 (within)	Temperature	Drug resistance, human adaptation, cattle health traits	[171–173]	RAN
5	6,551,121	E2F7 (68.592 kb)		Body weight, bone density (human)	[174,175]	RAN
6	57,392,860	TBC1D1 (within)		Selection (cattle, chickens), body size (chickens, mice), immune traits (lymphocytes, etc.), obesity (humans)	[176–182]	RAN
7	106,527,874	EFNA5 (within)	Precipitation			RAN
10	69,692,497	AP5M1 (within)		Selection (humans)	[183]	RAN
10	84,307,302	DPF3 (within)	Precipitation			RAN
15	2,232,970	GRIA4 (within)	Elevation	Cold tolerance (cattle)	[184]	RAN
15	71,402,156	LRRC4C (within)		Altitude adaptation (humans)	[185]	RAN

22	4,815,225	RBMS3 (156.7 kb)	Precipitation	Tropical adaptation (humans), Sweep region (humans), Pleiotropic QTL (cattle)	[17,186,187]	RAN
22	5,297,061	GADL1 (within)		Associated with climate variables in Mediterranean cattle, blood metabolites, human adaptation	[188–190]	RAN
24	10,628,280	CDH7 (150.58 kb)				RAN
24	35,718,635	ADCYAP1 (14.3 kb)		Circadian rhythm (birds), chronotype (humans)	[191–193]	RAN
25	35,085,041	CUX1 (67.8 kb)		Hair phenotypes (goats, mice)	[100,101]	RAN
25	39,464,325	LOC101904513	Precipitation	ncRNA		RAN
28	28,979,090	PLA2G12B (16.0 kb)	Temperature and Elevation	Local adaptation (humans)	[194,195]	RAN
29	44,555,972	BBS1 (within)		Obesity, energy homeostasis (human)	[107,108,196]	RAN
1	26,621,249	ROBO1 (within)		Neuron development (dogs, cattle, pigs), selection signature (cattle), sporting dogs, temperature acclimation (pigs)	[109,155,169, 197]	SIM
1	134,777,473	CEP63 (within)		Height (human), thermal adaptation (fish)	[198,199]	SIM
3	2,966,062	UCK2 (18.65 kb)		Osmoregulation (fish), disease response (cattle)	[200,201]	SIM
8	64,286,414	ENSBTAG00000054262		lncRNA		SIM
10	16,366,200	KIF23 (24.83 kb)		Hepatic function (cattle)	[202,203]	SIM

17	51,685,973	LOC100847522 (88.07 kb)		ncRNA		SIM
20	54,365,387		Precipitation, Elevation			SIM
23	1,768,070	LOC782044	Precipitation, Elevation			SIM
26	34,125,126	NRAP (within)		Meat traits (cattle), under selection in Eastern Finncattle, neuron development	[204–206]	SIM
28	640,998	RHOU (56.34 kb)	Elevation			SIM
29	36,766,544	LOC112444895 (137.67 kb)	Precipitation, Elevation	ncRNA		SIM
2	116,117,760	SPHKAP (within)	Elevation	Insulin secretion, kidney disease susceptibility (human)	[207,208]	GEL
3	65,627,833	ADGRL4 (39.34 kb)	Temperature			GEL
4	35,457,854	SEMA3D (within)		Calving ease, neuron/axon guidance	[209–211]	GEL
4	95,515,907	LOC785077 (30.81 kb)	Precipitation			GEL
11	81,911,781	FAM49A (26.09 kb)	Temperature			GEL
14	73,663,377	CALB1 (within)	Elevation	Selection signature (pigs)	[212]	GEL
23	1,760,296	LOC782044 (17.22 kb)				GEL
25	663,051	LOC531296 (within)/MSLN	Temperature	Feed intake (cattle)	[149]	GEL

Table 3.18. Gene enrichment analysis of envGWAS candidate genes in Red Angus, Simmental, and Gelbvieh populations. Candidate genes were annotated genes < 10 kb to significant envGWAS SNPs (p-value $< 1\times10^{-5}$). A single gene list for each breed was generated using significant SNPs from all combinations of univariate/multivariate, continuous/discrete envGWAS. Significant (FDR-corrected p-values < 0.1) KEGG pathways and GO biological processes are reported, along with the associated genes.

https://drive.google.com/file/d/1ZMVEYhND9WmvvrBhbyK7IKI3-hU-Im8n/view?usp=sharing

[Table S17 in text] [Tab S17 in above linked spreadsheet]

Table 3.19. TissueEnrich analysis using envGWAS gene sets from Red Angus, Simmental, and Gelbvieh populations. TSEA from TissueEnrich software using Human Protein Atlas gene expression data. Enrichment analysis carried out for candidate genes within 10 kb of significant envGWAS SNPs. For each test, we report the number of tissue specific genes, their average fold change, and the FDR-corrected $\log_{10}$ p-value for tissue enriched expression.

https://drive.google.com/file/d/1ZMVEYhND9WmvvrBhbyK7IKI3-hU-Im8n/view?usp=sharing

[Table S18 in text] [Tab S18 in linked spreadsheet above]

Table 3.20. Tissue enrichment analysis results from envGWAS gene sets in Red Angus, Simmental, and Gelbvieh populations using the pSI R package and human GTEx expression data. Enrichment significance values for four specificity index thresholds (pSI) of 25 human tissues types. Each combination of stringency for enrichment (pSI) and tissue reports a p-value for Fisher's Exact Test and a Benjamini Hochberg corrected p-value reported in parentheses. Tissue-gene-set combinations that are significant (Benjamini-Hochberg p-value < 0.1) are highlighted in red, those that are suggestive (raw p-value < 0.1) are highlighted in green.

https://drive.google.com/file/d/1ZMVEYhND9WmvvrBhbyK7IKI3-hU-Im8n/view?usp=sharing

[Table S19 in text] [Tab S19 in above linked spreadsheet]

Table 3.21. Brain region and cell-type enrichment analysis results from envGWAS gene sets in Red Angus, Simmental, and Gelbvieh populations using the pSI R package with expression data from the Allen Brain Atlas. Enrichment significance values for four specificity index thresholds (pSI) of six brain regions and 35 brain cell types. Each combination of stringency for enrichment (pSI) and brain region/cell type reports a p-value for Fisher's Exact Test and a Benjamini Hochberg corrected p-value reported in parentheses. Combinations that are significant (Benjamini-Hochberg p-value < 0.1) are highlighted in red, those that are suggestive (raw p-value < 0.1) are highlighted in green.

https://drive.google.com/file/d/1ZMVEYhND9WmvvrBhbyK7IKI3-hU-Im8n/view?usp=sharing

[Table S20 in text] [Tab S20 in above linked spreadsheet]

Table 3.22. Cell-type specific expression of homologous *C. elegans* genes derived from envGWAS candidate gene lists of Red Angus, Simmental, and Gelbvieh populations. *C. elegans* gene homologs were generated from Ortholist2, requiring that genes be present in at least three data sources to be included in enrichment analysis. For each breed's gene list, we include a list of worm tissues with significant enrichment of listed genes (q-value < 0.1).

https://drive.google.com/file/d/1ZMVEYhND9WmvvrBhbyK7IKI3-hU-Im8n/view?usp=sharing

[Table S21 in text] [Tab S21 in above linked spreadsheet]

SUPPLEMENTARY TEXT

Motivation for mapping polygenic selection

Though the first cattle genotyping assay was developed just over a decade ago [80], influential individuals from deep within the pedigree have been genotyped, providing a temporal distribution of samples at least ten generations deep for most numerically large breeds. Under directional selection, alleles will be at significantly different frequencies in more recent generations compared with distant ones. This creates a statistical association between allele frequencies at a selected locus and an individual's generation number. With multiple generations sampled and genotyped, we can disentangle small shifts in allele frequency due to directional selection from stochastic small changes from drift. Historical selective sweeps on simple traits in many species, including cattle, has been successfully studied, but recent and ongoing selection for complex traits is more difficult to characterize. Most methods used to identify selection rely on allele frequency differences between diverged populations (e.g. F_{ST}, FLK, XP-CLR) [213–215], or on the disruption of normal LD patterns (iHS, EHH, etc.) [216–218]. In cattle, these methods have successfully identified genomic regions under selection that control a handful of Mendelian and simple traits like coat color, the absence of horns, or large-effect genes involved in domestication [18,219–221] [77,155,176][18,219–221].

However, the traits actively under the strongest selection pressure in cattle are highly complex. This means that genomic changes due to selection are more likely to be subtle shifts (polygenic selection) at many loci rather than selective sweeps. Cattle producers are selecting on various combinations of growth, maternal, and carcass traits, but the genomic changes that result from this selection are not well-understood. Numerous genome-wide association studies have been undertaken on individual traits, but do not say anything about the underlying genomic changes that populations are experiencing. Our method, Generation Proxy Selection Mapping (GPSM) searches for allele frequency changes in a trait-agnostic manner by identifying statistical associations with an individual's generation (or some proxy). Cattle registered in herdbooks have recorded birth dates, making it a logical generation proxy to fit as the dependent variable in a linear mixed model (LMM)[13]. Generally, registered cattle populations have high levels of relatedness. Using a linear mixed model approach explicitly controls for confounding family and population structure with a genomic relationship matrix [3]. This allows GPSM to detect recent and ongoing polygenic selection within homogeneous populations (like breeds or smaller related groups).

Motivation for mapping local adaptation

Beef cattle are one of the few livestock species in the United States whose production environments remain largely unaltered by humans. Cattle are distributed across almost every possible environment in the continental United States [222], and virtually none are raised in controlled confinement like pigs or chickens. Local adaptation and genotype-by-environment (GxE) interactions exist in closely related cattle populations [223,224]. Previous work has identified the presence of extensive GxE in beef cattle populations [29,30,32,33,225], but limited work exploring the genomic basis of local adaptation has occurred [226].

Local adaptation has driven the development of environmentally adapted breeds and populations of cattle around the world. Well-adapted animals better express their genetic potential in their local environment, and thus are more likely to be selected as parents than their poorly-adapted counterparts. This leads to changes in allele frequency at loci that modulate adaptation in this local population compared to the full population. In the past 30 years, the use of artificial insemination has increased in U.S. beef cattle populations. This technology has allowed elite germplasm to be propagated ubiquitously across environments and has led to dramatic increases in production efficiency. However, as a consequence has made populations more homogeneous. We are interested in identifying whether detectable allele frequency differences exist between environments, and how human-imposed selection and mating has changed the magnitude of these differences.

To identify genomic regions potentially contributing to local adaptation, we used continuous environmental variables as quantitative phenotypes or discrete ecoregions as case-control phenotypes in a linear mixed model framework. We refer to these approaches as “environmental genome-wide association studies”, or envGWAS. Using a genomic relationship matrix in a LMM allows us to control the high levels of relatedness between spatially close individuals, and more confidently identify true signatures of local adaptation. This method builds on the theory of the Bayenv approach from Coop et al. (2010) [36,94] that uses allele frequency correlations along environmental gradients to identify potential local adaptation. Using genome-wide LMMs allowed us to extend the Bayenv method to extremely large datasets (eventually hundreds of thousands of animals with millions of SNPs). A GWAS model in conjunction with a GRM controls for cryptic family and population structure that frequently obscures these studies.

We use two definitions of environmental dependent variables in our envGWAS studies. First, we directly fit 30-year normal values for temperature, precipitation, and elevation for the zip code attached to a genotyped animal. Additionally, we create 9 discrete statistically-derived ecoregions based on *K*-means clustering of these three environmental variables. Using discrete environments as dependent variables in a case-control manner allows us to capture unique combinations of these three measures of environment. These regions recapitulate known production environment differences for beef cattle. For instance, the "Fescue Belt", the region of the US where tall fescue (*Festuca arundinacea*) is the most common forage, is almost perfectly defined using only these three environmental variables. Fescue is a major source of forage, but harbors a fungal endophyte, *Epichloë coenophiala*, that produces ergot alkaloids that cause multiple physiological challenges for cattle that graze it [28]. Beyond fescue's endophytes, we know that local and regional parasite and pathogen loads differ [227], and we anticipate that using these discrete definitions of climate will serve as a proxy for these and similar environmental and ecological phenomena.

Simulations

To ensure that GPSM signal detected in our three real datasets was being driven by selection and not genetic drift, we performed two major sets of simulations, stochastic and gene drop. First, we performed a set of stochastic simulations to demonstrate how selection in the context of different effective population sizes, selection intensities, generational sampling, and genomic architectures produce GPSM signal. In each test case, using the same starting population, we perform parallel regimes of selection; one where the selection of parents is random, and the other where selection is based on an animal’s true breeding value. In the random selection regime, we expect drift would be the sole driver of allele frequency changes.

Second, we performed multi-locus gene dropping simulations with a full pedigree containing the 63,122 registered animals related to the 15,323 genotyped purebred Red Angus individuals to test whether drift in the context of biased sampling could produce a GPSM signal. Additionally, we used these same gene drop experiments, but with continuous environmental phenotypes to test whether genetic drift can create significant envGWAS associations (. However, this simulation cannot distinguish between drift, pedigree structure, and founder effects. This is a function of only elite, influential sires having cryopreserved semen available for genotyping. Further, early in the implementation of genomic prediction in the beef industry, high ranking animals were preferentially genotyped.

All simulations were performed in AlphaSimR [228]. Stochastic simulations were performed in 10 replicate sets using 10 sets of founder haplotypes as starting points. We generated founder haplotypes using the AlphaSimR wrapper around MaCS [229]. Using an approximation of the demographic history of cattle, we simulated 10 chromosomes with 20,000 segregating sites each for 2,000 founder individuals (1,000 males and 1,000 females). This resulted in a starting effective population size (N_e) of approximately 100, similar to estimates of U.S. beef cattle populations [13]. To test other N_e, we simulated populations with effective population sizes of 50 and 250. Based on the chosen genomic architecture, 1000, 500, or 200 purely additive QTL were randomly assigned to segregating sites. Effect sizes for simulated QTL were drawn from either a normal (mean = 0, variance = 1) or a gamma (shape = 0.42) distribution [230]. Prior to the two divergent selection regimes, we performed five generations of burn-in selection to establish LD in our populations.

After burn-in (generation 0), we performed selection of parents for the next generation in two parallel manners: randomly or truncation selection on true breeding value. In each scenario, we held the effective population size by selecting appropriate numbers of males and females to be parents each generation. Selection intensity was altered by increasing or decreasing the number of crosses performed (1000, 2000, 4000, 8000). We also varied the number of generations of selection post-burn-in (20, 10, and 5 generations). For each scenario, we extracted 10,000 total simulated individuals for analysis in GPSM. To test the effects of uneven generation sampling that we see in real data, we performed two different strategies for sampling simulated genomes. In one case, we sample an equal number of individuals each generation. In the other, we sample more animals from the most recent generations. The number of sampled individuals is based on a negative exponential distribution that approximates the of ages observed in our real datasets (**Fig S1**). Sampled individuals were chosen at random, and were not more or less likely to become parents in the next generation. In addition to sampling genotypes each generation, we calculated the allele frequency of simulated QTL each generation to track observed allele frequency changes over the course of selection. This process was performed in replicates of 10 for each scenario, allowing us to calculate descriptive statistics and compare GPSM’s performance across scenarios.

After sampling genotypes, we created a standardized genomic relationship matrix (GRM) in GEMMA (v0.98.1) with all SNPs that had a MAF > 0.01. Using GEMMA, we fit the individuals' true generation number as the dependent variable in a genome-wide linear mixed model. Outputs from GPSM were read, manipulated, and plotted in R using multiple tidyverse packages [231].

Stochastic simulations under multiple selection intensities, time periods, and trait architectures showed consistently that GPSM is able to map polygenic selection (**Table S2**). Across all simulated scenarios and architectures, we identify an average of 38.5 simulated QTL (min 5.2, max 64.1) with GPSM. In many of these scenarios, we observe that significant hits are not the largest effect simulated QTL, but the loci that have undergone the greatest allele frequency shifts over the course of genotype sampling. In many cases the largest effect QTL had been fixed in the population during burn-ins, making their detection by GPSM impossible. GPSM's ability to detect selection relies on the selection occurring during the period of sampling.

Simulations also suggest that GPSM is able to effectively distinguish allele frequency changes due to selection from those associated with drift. Across 10 replicates of 36 scenarios of random selection, we detect an average of only 0.7 GPSM false positive SNPs (sd = 0.85). These rare false positives do not appear to be driven by changes in any single component of the simulations.

We simulated haplotypes in MaCS for our 5,223 founder individuals. Founder haplotypes spanned 10 chromosomes, each with 20,000 segregating sites for a total of 200,000 SNPs. These founder haplotypes were then randomly dropped through the Red

Angus pedigree, restricted to ancestors of genotyped individuals, in a Mendelian fashion, with recombinations occurring at a rate of one crossover per Mb.

Gene dropping simulations using the Red Angus pedigree generated an average of 6.7 (sd = 4.14) significant GPSM SNPs per 200K markers tested, equating to what would be ~27 markers in our 850K data (**Table S1**). These SNPs represent, on average 3.9 (sd = 1.37) genomic regions per 200K SNPs. This means that pedigree structure is responsible for a small portion of the significant SNPs that we detect in our real datasets. That said, the pedigree structure of these cattle populations is largely created through directional selection over time, meaning that we cannot completely disentangle the drift and selection components here. Further, these 27 markers would make up ~ 10% of the total GPSM SNPs identified in the Red Angus dataset, suggesting that selection pressure independent of nonrandom sampling is generating the majority of observed GPSM signals.

GPSM detects signatures of polygenic selection across and within three populations of U.S. Beef Cattle

In addition to the locus undergoing a massive allele frequency shift on BTA28 (lead SNP *rs1762920*), we identify 6 other loci under selection in all three populations. (**Table S7**). These shared GPSM signals suggest that not only are there similar selection pressures, but common genomic architectures under selection in these three populations . While we identify significantly more population-specific GPSM signatures, shared signatures are of interest as they are likely serving an important role in all breeds of beef cattle. We discuss these loci and their potential functions in beef cattle that are driving allele frequency changes.

The largest genomic region detected by GPSM lies at the end of BTA1 (157.5 Mb - 158.5 Mb). The lead SNP in this peak (*rs1755753*) lies within the long non-coding RNA (lncRNA) *LOC112448253*. This lncRNA has not been previously associated with any traits or function in cattle. This region contains dozens of other potential candidate genes. This selected region is also immediately upstream of *PRDM9*, a modulator of recombination in most mammalian species, including cattle [87,232,233]. While no variants within PRDM9 reach genome-wide significance, variants ~10.7 kb from the TSS are responding to selection. Selection on *PRDM9* could increase average recombination rate or allow novel motif binding in order to create novel favorable haplotype combinations [234].

We identified six other common genomic regions under strong selection in all three populations, encompassing 106 statistically significant markers. Another 90 SNPs overlap in at least two populations, corresponding to 15 additional genomic regions under selection (**Table S7**). All three populations have been selected for increased growth traits over the last 50 years [84], and GPSM identifies two genomic regions that have been previously associated with feed efficiency and growth traits. The common peak on BTA12 (lead SNP *rs1389713*) is ~182 kb upstream of the *DACH1* (Dachshund homolog 1), a transcription factor associated with post-weaning gain, various indicators of feed efficiency [139,235], and backfat thickness [236] in cattle. The shared GPSM peak on BTA14 resides near a known QTL for post-weaning gain near the gene *LRP12* (LDL receptor related protein 12) [53].

In addition to selection on loci that appear directly involved in growth and efficiency, we identify multiple selection targets likely involved in aspects of immune function. A shared significant peak on BTA2 (lead SNP *rs1080110*) resides within the *ARHGAP15* gene (Rho GTPase Activating Protein 15) that is essential for Trypanosomiasis resistance in African cattle populations [136–138]. Though Trypanosomiasis and other tsetse fly-transmitted diseases are restricted to Africa, genetic tolerance to similar immune disturbances may account for the positive selection observed in these three American cattle populations. Variants within *ADORA1* (Adenosine A1 Receptor) are also detected as being under selection by GPSM. In cattle, *ADORA1* plays a role in the activation of polymorphonuclear neutrophilic leukocytes, which are important for peripartial immune responses in cattle [141], and likely play roles in other immune functions. *ADORA1* and other purinergic receptors play an important role in bone metabolism [237] and likely growth in cattle. This signature within *ADORA1* also spans a potentially regulatory region for *MYBPH*, an important gene in muscle formation and development, and another potential target of selection [142]. In this case and others, we identify multiple logical candidate genes within regions undergoing selection. We report a complete listing of significant SNPs and candidate genes from GPSM analyses of each population in (**Table S5**).

We observe 22 genomic regions that are changing in frequency in at least two of our datasets. In addition to these shared loci, we identify evidence of shared networks and genetic architectures under selection (**Table S8**). To discern common biological pathways and processes under selection across multiple populations, we identified genes within 10 kb of SNPs identified by GPSM in at least two populations and performed a gene enrichment analysis in ClueGO. Using the 46 genes residing in or near the 29 shared GPSM signatures identified by at least two datasets identified multiple biological processes undergoing selection. The most significant pathways involved G-protein coupled signaling (Benjamini–Hochberg adjusted p-value = 0.038) and purinergic receptor signaling pathways (Benjamini–Hochberg adjusted p-value = 0.034, associated genes *ADORA1* and *P2RY8*). Purinergic receptors have been identified as important drivers of immune responses in cattle [141]. Selection on immune pathways is likely driven by the increased production efficiency of healthy calves [85,238]. Shared selection on *SLC2A5* and *BIRC5* point towards biological pathways involved in sensing carbohydrate, hexose, and monosaccharide stimuli (Benjamini-Hochberg adjusted p-value = 0.01). We also expect that an enhanced metabolic response to carbohydrates would result in increased animal efficiency. Finally, genes involved in the regulation of arterial blood pressure (Benjamini–Hochberg adjusted p-value = 0.045: *ADORA1*, *SLC2A5*) made up the lone other significant gene class under selection across all three populations. Gene enrichment analysis within populations also identified population-specific pathways and processes under selection. In Simmental cattle we detect 23 GPSM candidate genes involved in olfactory transduction. While not related directly to growth traits, olfactory receptors have been identified as selection targets in many mammalian species, including cattle [156,176,239]. This rapidly-evolving class of genes is also associated with growth and carcass traits, suggesting a wide range of functions that are not limited to detecting smell [53,240].

Principal Component Analysis

While we expect modest allele frequency differences in these populations, we assume that they are largely panmictic due to extensive gene flow via artificial insemination. We performed a principal component analysis (PCA) in each dataset to make sure that existing population structure is not driven by ecoregion-of-origin. We see minimal evidence that ecoregion-of-origin is a major source of structure in these populations (**Fig S5**). This suggests that genetic variation is significantly greater within ecoregions than between. This is in contrast with the estimates of genetic variation explained by environment (**Table S12-S13**). Taken together, we observe that these populations are largely panmictic, but still have genetic associations with the environment. Some of these environmental associations are driven by family relatedness, but others are strong even after accounting for familial relationships with a genomic relationship matrix (envGWAS signatures).

envGWAS identifies adaptive pathways and processes.

Local adaptation is likely highly complex and controlled by many areas of the genome. Though we detected minimal overlap in candidate genes across datasets, we identified multiple conserved biological processes and pathways that appear to play roles in local adaptation across populations.

Though there was minimal candidate gene overlap between continuous and discrete envGWAS (19 of 187 total envGWAS candidate genes in Red Angus), we identified many shared pathways and gene ontologies. In most cases where we observed GO or pathway overlap, statistical significance was greater for the discrete envGWAS analysis, simply due to the difference in number of provided genes (168 vs 38). Most of the shared terms and pathways were driven by the genes *BAD*, *EFNA5*, *LRRC4C*, *MARK2*, *PLCB3*, and *PRKG2* detected in both analyses. In these overlapping terms, additional genes from the discrete zone envGWAS further supplemented the identified terms.

We performed gene enrichment analyses with gene lists from all univariate and multivariate, discrete and continuous envGWAS analysis in each population. This allowed sufficiently large gene lists to identify potentially adaptive pathways and processes. We were particularly interested in identifying shared pathways and processes between populations since the number of shared genomic regions was low. Across all three populations, we consistently identified the “axon guidance” pathway, and numerous GO terms relating to axon development and guidance under region-specific selection. Ai et al. (2015) [109] suggested that axon development and migration in the central nervous system is essential for the maintenance of homeostatic temperatures by modulating heat loss or production [241]. The direction and organization of axons is an essential component of the olfactory system which is frequently implicated in environmental adaptation through the recognition of local environmental cues [17,242]. Other pathways identified across all three datasets include “cholinergic synapse”, “glutamatergic synapse”, and “platelet activation”. Cholinergic signaling drives cutaneous vascular responses to heat stress in humans [243–245]. Additionally, cholinergic receptors act as the major neural driver of sweating in humans [246]. Glutamatergic synapses are involved in neural vasoconstriction [247,248]. “Retrograde endocannabinoid signaling”, “dopaminergic synapse”, “GABAergic synapse”, and “serotonergic synapse” pathways are also significantly enriched by envGWAS candidate genes. Nearly all of these important neural signaling pathways are also enriched in a series of gene-by-environment interaction GWAS for birth weight, weaning weight, and yearling weight in Simmental cattle by [33]. Taken together, these results suggest that pathways involved in neuron development and neurotransmission are essential components of local adaptation in cattle. Temperature homeostasis is largely controlled by the central nervous system [110], making environment-specific selection on these pathways an efficient way for populations to adapt.

Other pathways identified by envGWAS appear to play important roles in vasodilation and vasoconstriction. Relaxin signaling was identified in both Red Angus and Simmental populations as a locally adaptive pathway under selection. Relaxin, initially identified as a pregnancy-related hormone, is an important modulator of vasodilation [249]. In Simmental this pathway association originates from local adaptation signatures near the genes *COL1A1*, *MAPK10*, and *PRKACA* identified both in our discrete multivariate envGWAS and in the Desert ecoregion univariate analysis. Relaxin signaling was also identified in Red Angus, but with four entirely different genes (*LOC529425*, *PLCB3*, *PRKACB*, *VEGFB*). While all four of the Red Angus genes were identified in multivariate analyses, we identify two of them in the Desert ecoregion univariate envGWAS. Vasodilation is an essential component of physiological temperature adaptation in cattle and other species [111,250,251]. The ability to mount a physiological response to heat stress has a direct impact on cattle performance. Heat stressed cattle have decreased feed intake, slower growth rates, and decreased fertility [252]. The Renin secretion pathway, which is also directly involved in vasoconstriction was identified in Red Angus (*ADRB1*, *PLCB3*, *PRKACB, PRKG2*), and has also been previously implicated in physiological responses to heat stress in cattle [253]. In each population we identify multiple biological processes related to the regulation of insulin secretion. Insulin secretion is elevated in heat stressed cattle [254] and pigs [255], suggesting that it plays a role in metabolism and thermoregulation. Insulin secretion in response to the presence of glucose may also be related to different diets and forage availability along these continuous environmental gradients [256].

Other pathways identified by envGWAS candidate genes from Simmental point towards the immune system’s role in local adaptation. “Th1 and Th2 cell differentiation” and “Th17 cell differentiation” were significant in a KEGG pathway analysis. The development of these cell types is essential for adaptive immune responses [257,258]. These signals were driven in part by region-specific allele frequency differences in or near *MAPK10* an innate immunity gene identified in human studies as an adaptive target of selection [259]. We also identify multiple cardiac-related pathways from Simmental envGWAS genes (“Dilated cardiomyopathy”, “Arrhythmogenic right ventricular cardiomyopathy”, “Hypertrophic cardiomyopathy”) driven by a pair of related genes *SGCA* and *SGCD*. Like other circulatory-related pathways, the alleles driving this signal were identified in both multivariate and Desert ecoregion envGWAS. We expect that these cardiac-related pathways, like renin secretion and relaxin signaling affect the efficiency of circulation and improved temperature homeostasis when exposed to heat or cold stress. Selection on cardiovascular function is also likely a central component of adaptation to high altitude [260].

Tissue Set Enrichment Analysis

To further disentangle the biological basis of GPSM and envGWAS adaptive signatures, we performed a series of Tissue Set Enrichment Analyses (TSEA) based on gene expression data from humans and worms (*C. elegans*). These analyses identified tissues in which our envGWAS candidate genes were preferentially expressed. Our candidate gene lists for each population consisted of annotated cattle genes within 10 kb of a significant GPSM SNPs or SNPs identified in any of our multivariate or univariate envGWAS analysis. In using expression data from other species, non-orthologous cattle genes are not included in enrichment analyses.

Using GPSM candidate genes in TSEA identified enriched tissues that correspond to population-specific production traits known to be under selection. Using gene expression measures from the GTEx pilot dataset [131] in the pSI R package [132] we identify suggestive enriched expression in various reproductive-related tissues from the Red Angus GPSM candidate gene set. We observe suggestive enrichments ($p < 0.1$) for human breast, ovary, pituitary, and uterus tissues (**Table S9-S10**) among others. An analysis of this gene list with gene expression data from the Human Protein Atlas also identified significantly enriched expression in cervix, uterus, and brain tissues (FDR-corrected p-value < 0.1) and suggestive expression fold change in the ovary. These results provide further evidence that selection on fertility and reproductive traits have been ongoing in the Red Angus population over the last ~10 generations (Red Angus Association of America EPD trends https://redangus.org/genetics/epd-trends/). Further, using Gelbvieh GPSM candidate gene sets, we identify enriched expression in numerous tissues including skeletal muscle, nerve, thyroid and adipose tissue. These enriched tissues align with known ongoing selection for increased growth and carcass quality.

Despite the numerous enriched pathways and processes from the Simmental GPSM gene lists, we identified minimal tissue-specific expression. We did not identify significant or suggestive tissue enrichments in the Simmental GPSM gene set using the Human Protein Atlas in TissueEnrich. Uterine tissue was the only tissue showing suggestive enrichment. Pathway analyses of envGWAS candidate gene sets in all three populations pointed towards a role in neural development and signaling in modulating adaptation. We used TSEA in humans and *C. elegans* to provide further evidence of brain and nervous system tissues involvement in environmental adaptation. Using *C. elegans* tissues allowed us to refine the expression of conserved genes to individual neuron resolution. Using gene expression data from the Human Protein Atlas in the TissueEnrich software, we identify the cerebral cortex to be the lone significant tissue among our candidate gene sets from Red Angus and Gelbvieh envGWAS analyses (**Table S18-S19**). These results agreed with TSEA using GTEx data. Despite identifying similar neural pathways in the Simmental population, we did not observe enriched expression in brain tissue in either human TSEA.

To further probe the specific brain regions expressing envGWAS candidate genes, we performed a brain-specific expression enrichment analysis using the pSI tool with human brain expression data from BrainSpan and the Allen Brain Atlas [261]. Complete results are reported in **Supplementary Tables 20-21.** The only brain region significantly enriched for envGWAS candidate genes was the cortex in Simmental ($p = 1.525\text{x}10^{-4}$). Interestingly, Simmental was the only population in which Brain tissue did not show enriched expression in the GTEx data. The envGWAS candidate genes from Red Angus and Gelbvieh showed suggestive enrichments in expression in the Cerebellum ($p = 0.096$)

and Striatum, respectively (p = 0.059). While some suggestive cell-type specific expression differences existed with each gene list, we did not observe any tissues with conserved expression across populations.

GPSM meta-analysis identifies loci responding to selection in an ecoregion-specific manner.

We use a combination of our discrete ecoregions and GPSM to identify region-specific signatures of selection. This approach draws on the GxE interpretation of metaGWAS analyses from Kang et al. (2014) [262], where a heterogeneity of effect sizes at a locus between “treatments”, or in our case ecoregions, is indicative of an environment-specific association. This analysis identified variants undergoing region-specific changes in allele frequency. If selection on locally adaptive variants is ongoing, we would expect this analysis to identify variants actively diverging from the population mean in a particular ecoregon. The converse, an erosion of allele frequency differences back to the population mean, would suggest that selection for locally adaptive variants is not strong enough to overcome the influx of alleles from the outside population via artificial insemination (**Fig S 9-11**). We performed a meta-analysis in METASOFT [124] by incorporating individual GPSM analyses for each ecoregion with more than 1,000 sampled individuals. For each marker tested in the meta-analysis, a Cochran’s Q value is assigned based on the heterogeneity of effects between ecoregion. Variants with high Q values are experiencing directional selection in one or more regions and little to no directional selection in other regions.. The test also generates within-region GPSM p-values, an across-study p-value optimized for detecting associations in contexts of increased heterogeneity [124], and within-region m-values, the posterior probability that an effect exists within that region. We perform this test for all three populations using ecoregion cohorts with > 1,000 animals to ensure adequate sample sizes in GPSM.

CHAPTER 4

UNCOVERING THE ARCHITECTURE OF SELECTION IN *BOS TAURUS* BEEF CATTLE

Troy N. Rowan[1], Robert D. Schnabel[1,2], Jared E. Decker[1,2]*

[1] Division of Animal Sciences, University of Missouri, Columbia, MO 65211 USA

[2] Institute for Data Science and Informatics, University of Missouri, Columbia, MO 65211 USA

*Corresponding Author

Abstract

Mapping polygenic selection on complex traits is difficult because it does not leave clear signatures on the genome like a selective sweep. In cattle populations with temporally-stratified genotypes, we can use genome-wide linear mixed models to identify allelic associations with an individual's generation (or some proxy) with the Generation Proxy Selection Mapping (GPSM) method to identify those variants that have experienced the greatest allele frequency changes. Here, we use GPSM on two large datasets of beef cattle to detect associations between an animal's generation and 11 million SNPs using imputed genotypes. Using these datasets with high power and base-pair mapping resolution, GPSM detected a total of 1,015 unique loci actively under selection in the Simmental and Red Angus breeds. We observed that GPSM has a high power to detect selection in the very recent past (< 10 years), even when allele frequency changes are relatively small. Variants identified by GPSM reside in genomic regions associated with known breed characteristics and selection goals, such as fertility and maternal ability in Red Angus, and carcass and coat color in Simmental. Greater than 60% of the selected loci reside in or near (<50 kb) annotated genes. Many more selected loci overlap known epigenetic marks or genomic regions that are likely to be functional. Selected loci have little overlap with historically selected loci identified by selective sweep mapping methods, making it a complementary approach to sweep detection methods when temporal genotype data are available. We demonstrate that GPSM is a powerful strategy for understanding the genetic architectures on which polygenic selection acts.

Introduction

Since their initial domestication in the Fertile Crescent ~10,500 years ago [263] cattle have been exposed to intense selection for increased tameness, production, and fecundity leading to substantial changes in these phenotypes. While selection occurs on phenotypes, and more recently on estimated breeding values, the phenotypic changes generated in a population are due to changes in the trait's underlying genotype frequencies. Occasionally, selection on beneficial mutations of very large phenotypic effect can lead to very rapid changes in allele frequency at a locus. The resulting "selective sweep" not only increases the frequency of the beneficial variant, but also reduces genetic variation in the genomic region surrounding the selected locus [16]. In cattle populations, multiple sweeps have been mapped to genomic regions, many of which are related to simple Mendelian traits such as polledness (the absence of horns) [219] and coat color [155] or large-effect quantitative trait loci (QTL) influencing production traits [77].

While sweeps at Mendelian loci have played an important role in the domestication and improvement processes of cattle, it is becoming increasingly apparent that the majority of both historical and ongoing selection is on highly complex traits [264,265]. Complex traits are controlled by many mutations of relatively small effect spread throughout the genome [266]. Under complex trait architectures, selection can generate substantial changes to a phenotype without necessarily generating large allele frequency shifts [21]. These modest directional allele frequency changes at selected loci make mapping selection on polygenic traits over short timescales is difficult. However, by leveraging large commercially-generated datasets that include influential founder individuals, cattle populations offer intriguing opportunities for mapping the loci exposed to polygenic selection over time [50]. The Generation Proxy Selection Mapping (GPSM) method uses a proxy for the number of meioses that separate an individual from the beginning of the pedigree as the dependent variable in a genome-wide linear mixed model to detect significant associations between generation and allele frequency [13,25]. When applied to cattle populations, it is effective at detecting subtle ongoing shifts in allele frequency across the genome of multiple cattle populations [14].

Genome-wide association studies have motivated extensive work attempting to dissect the genetic architecture of complex traits in many species [267]. We expect that selection on these complex traits occurs on similarly unique architectures [21]. Here, we expand on previous work exploring the selection landscape in cattle with one of the largest non-human selection mapping datasets explored to date. By using sequence-imputed genotypes for over 124,000 individuals in two beef cattle populations, we identify subtle ongoing shifts in allele frequency due to selection at the base-pair level of resolution. With these high-resolution data, we can discern the underlying functional genomics on which selection acts in cattle populations. Additionally, we use subsets of these data to explore recent and ongoing selective sweeps using both haplotype and site frequency spectrum (SFS)-based approaches. This combination of selection mapping approaches provides a detailed report of the genetic architectures on which strong directional selection acts in cattle populations, and a blueprint for leveraging temporally-distributed genotypes to understand the selection architectures of other species.

Results

Quantitative Genetic Signals of Polygenic Selection

In each population, we used genomic restricted maximum likelihood (GREML) [268] to estimate the proportion of variance explained (PVE) by 811,967 imputed SNPs in various subsets of our Simmental (SIM) and Red Angus (RAN) datasets (**Supplementary Tables 1 & 2**). Using an individual's date of birth as a generation proxy, we estimated the PVE to be 0.523 (se = 0.007) and 0.619 (SE = 0.005) in RAN (n = 46,454) and SIM (n = 78,787), respectively). This suggests that even while controlling for relatedness, selection makes the most recently born individuals more similar genomically than those born at early time points. Due to the non-normality of sampled birth dates in both datasets (**Supplementary Figure 1**), we observe a large divergence from expectation in individual breeding values and residuals in GREML analyses, particularly for the earliest born individuals (**Supplementary Figure 2**). To help normalize residuals and potentially boost our power to detect selected variants, we performed a series of transformations to birth date (**Supplementary Tables 1 & 2, Supplementary Figure 3**). When using log-transformed birth date as the dependent variable in Red Angus, PVE increased to 0.657 (se = 0.006). Using log-transformed birth date in Simmental decreased PVE to 0.600 (se = 0.005). Other transformations and their impacts on PVE are reported in **Supplementary Tables 1 & 2**.

We explored the effects of various statistical transformations on our ability to detect GPSM signatures in Red Angus using the 811K SNP dataset. The number of significant SNPs identified using an identical dataset, but with different statistical transformations performed on the generation proxy provides a measure of statistical power. In Red Angus, we observed an increase in the number of variants detected at multiple significance thresholds (nominal: $p < 10^{-5}$, Bonferroni: $p < 7.55 \times 10^{-7}$, and FDR-corrected q-values < 0.1 and < 0.05). At all significance thresholds, log-transformed birth date detected between 61% to 79% more significant SNPs and at least 12 additional significant loci (**Supplementary Table 3, Supplementary Figure 4**). A GPSM analysis using birth dates of animals born since 2012 detected almost all of the same loci one that used log-transformed birth date (**Supplementary Figure 4**). The same transformations applied to the Simmental dataset led to a reduction in the number of identified SNPs and loci, compared to using raw birth date (**Supplementary Table 4**). Interestingly when analyzing data only for animals born since 2012, GPSM identified almost all of the same loci as did the log-transformed birth date GPSM on the full Red Angus dataset.

While the Red Angus dataset was composed of almost entirely purebred individuals, the Simmental dataset contained large numbers of crossbred animals. The numbers of animals with non-Simmental ancestry that have been genotyped by the breed association have significantly increased in recent generations (**Supplementary Figure 5**). Consequently, we divided the Simmental dataset into subsets based on pedigree-reported ancestry and/or birth date. This allowed us to examine the selection that is occurring within different subsets of the population. The PVE for birth date remained moderately high, ranging from 0.619 (SE = 0.005) for all animals with > 5% SIM ancestry to 0.436 (SE = 0.021) for animals born before 2008. A complete accounting of variance components estimated from these subsets is in **Supplementary Table 2**.

Generation Proxy Selection Mapping (Red Angus)

Based on results from the 811K GPSM analyses described above, we performed three separate sequence-level analyses (11,759,568 imputed SNPs) in Red Angus using birth date as the dependent variable for all individuals, for animals born before 2012, and animals born after 2012, referred to hereafter as full Red Angus, old Red Angus, and young Red Angus, respectively.

GPSM identified 2,914 SNPs, 9,065, and 0 SNPs (Bonferroni-corrected threshold $p < 4.29 \times 10^{-9}$) significantly associated with birth date in the full and young Red Angus datasets, respectively. (**Figure 1**). A less stringent p-value threshold of 5×10^{-8} identified 3,617, 10,939 and 0 SNPs associated with birth date in the full, young, and old Red Angus datasets, respectively. The old dataset, which contained only 1,984 individuals, was likely underpowered to detect signals of selection using GPSM. There were 3,240 SNPs identified in the full dataset that were also significant in the young Red Angus data, a near-complete overlap. A stepwise conditional & joint analysis (COJO) [269] of these results further refined lead SNPs in significant loci and identified additional independent associations (COJO $p < 5\times10^{-8}$). COJO identified 248 independent associations with birth date in the full dataset and 417 in the young dataset. Despite these datasets being largely comprised of the same individuals and sharing a large proportion of common GPSM SNPs, only 25 SNPs identified by COJO were shared between both datasets.

Using COJO SNPs from the sequence-level GPSM analysis, we annotated genes, known quantitative trait loci (QTL), and other genomic features to help understand the biological pathways and phenotypes that selection targets in these populations. In all datasets, the majority of COJO SNPs resided within, or adjacent to (within 50 kb) annotated genes (**Table 1**). Depending on the dataset, 46-56% of these genes with a clear positional candidate gene resided in regions immediately upstream or downstream of transcription start sites, insinuating that a high proportion of selective pressure is on *cis*-regulatory regions of the genome. A complete accounting of positional candidate genes for COJO SNPs in the full and young Red Angus datasets is provided in **Supplementary Tables 5 & 6**, respectively.

In many cases, GPSM in young Red Angus animals detects novel signatures of recent ongoing selection that are not significant in the full dataset. Chromosome 2 offers an interesting example of these differences. In addition to the two major peaks identified by both the full and young Red Angus datasets, at least eight additional major peaks are identified in the young data, accounting for 48 unique COJO associations (**Figure 2 A & B**). The strongest unique associations identified in the young Red Angus dataset reside within the gene *ARHGAP15*. This association contains seven unique COJO SNPs within the gene. *ARHGAP15* is a major gene involved in trypanotolerance in African cattle [136,138], and likely has wider effects on immune function in worldwide cattle populations. Further, *ARHGAP15* is almost exclusively expressed in immune tissues [270].

Using previously identified and annotated QTL for cattle, allowed us to interpret the biological and production impact of selection decisions reflected in the GPSM results. A QTL-enrichment analysis identified 12 QTL classes significantly enriched near GPSM COJO associations (**Supplementary Tables 7 & 8**). The largest QTL enrichment in Red Angus was for metabolic body weight (FDR-corrected $p = 4.13 \times 10^{-82}$) , driven primarily by GPSM signals on chromosomes 14 (23.3 Mb) and 6 (38.5 Mb). Other enriched QTL pointed towards ongoing selection for increased fertility and calving ease, known selection goals in the breed (https://redangus.org/about-red-angus/history/). Calving ease, birth weight, stillbirth (maternal & direct), and sexual precocity QTL were each enriched among QTL tagged by GPSM COJO SNPs across the genome (FDR-corrected $p < 7.056 \times 10^{-7}$).

We used a meta-assembly of selection signatures in cattle by Randhawa et al. (2016) [24] to quantify the proportion of genomic regions identified by GPSM that had previously been identified as harboring selection signatures in cattle. Despite the different underlying strengths and goals of traditional selection mapping and GPSM, we found that GPSM identified multiple regions of the genome that had previously been identified as under selection in cattle populations using a variety of methods. Of the 208 positional candidate genes (<50 kb from GPSM COJO hit) in our full Red Angus dataset, 105 were also present in the selected loci in the selection signature meta-assembly. In the young Red Angus dataset, 142 of the total 272 GPSM identified genes were also identified by Randhawa et al. (2016).

Generation Proxy Selection Mapping (Simmental)

We individually analyzed the purebred Simmentals (100% Simmental ancestry) in our dataset to identify ongoing selection restricted to the breed. We performed an additional sequence-level GPSM on the full Simmental dataset (all animals with > 5% reported Simmental ancestry), representing selection within the entire herdbook, including the full spectrum of admixed animals.

GPSM analysis of purebred Simmental animals (n = 13,379) identified 513 significant SNPs (**Figure 3 A & B**) (Bonferroni-corrected p-value threshold p < 4.29 x

10^{-9}), led by a strong signal on chromosome 5, centered at a locus containing *PMEL* and *ERBB3*, genes known to control coat color in Fleckvieh populations (European Simmental) [82]. This locus, coupled with a GPSM signature immediately upstream of *KIT [271]*, suggest that the strongest selection pressures in the purebred American Simmental population have been on changing coat color and external appearance, making them appear less like European Simmental and more like American Angus. The next most significant locus resides in a cluster of olfactory receptors on chromosome 28. Another strong signal exists on chromosome 15 near beta-carotene oxygenase 2 (*BCO2*) and interleukin-18 (*IL-18*), a locus likely involved in immune functions [272]. Within these 513 genome-wide significant SNPs, COJO identified 33 independent associations, 13 of which were located at the center of chromosome 5 in the *PMEL/ERBB3* locus.

A GPSM analysis of all registered Simmental animals with at least 5% Simmental pedigree ancestry (n = 78,787) identified 1,008 genome-wide significant SNPs (Bonferroni-corrected $p < 2.29 \times 10^{-9}$) (**Figure 3 A & B**). A COJO analysis found 334 independently-associated SNPs (COJO $p < 5 \times 10^{-8}$), two of which were identified in the purebred analysis (14:28004:G:T, 8:61247:T:C). An additional 4 loci were shared between datasets, despite different COJO SNPs being identified within these loci. These include the same significant coat color-associated regions on chromosomes 5 and 6 and olfactory receptor cluster on chromosome 28 mentioned above.

By subsetting these data, we identified recent Simmental-specific selection in the purebred dataset and contrasted it with signatures identified in the full dataset where introgression from other breeds is also responsible for changing allele frequencies. In the full dataset, the most significant COJO SNP (9:66692267:C:T, COJO $p < 10^{-310}$) is located immediately upstream of *PTPRK*, a gene associated with marbling and tenderness measures in beef cattle [273,274]. Another strong GPSM signature (9:75790346, COJO p $< 10^{-310}$) was found within the *TNFAIP3* gene, which has been implicated in multiple studies of bovine and human immune function [275–278]. We also identified strong selection on variants (4:94308044:A:G, COJO $p = 2.17 \times 10^{-170}$) in the imprinted gene *COPG2* [279]. As in Red Angus, the majority of GPSM COJO SNPs were either in, or proximal to, (< 50 kb from) genes (**Table 1**). In both the full and purebred datasets 30% of COJO SNPs resided within genes. An additional 35% and 40% of COJO SNPs resided in close proximity to genes in the full and purebred datasets, respectively. A complete accounting of Simmental GPSM COJO detected SNPs and their positional candidate genes are in **Supplementary Tables 9 & 10.**

In the purebred Simmental dataset, two of the five significant QTL enrichments were involved with the appearance-based traits eye pigmentation (FDR-corrected $p = 10^{-4}$) and coat color (FDR-corrected $p = 5.47 \times 10^{-3}$), largely driven by the selected loci on chromosomes 5 and 6 (**Supplementary Table 11**). While the same significant loci were also identified in the full Simmental GPSM analysis, the enriched QTL classes under selection in the full dataset included carcass (longissimus muscle area , carcass weight, and meat color), production (metabolic body weight, average daily gain, dry matter intake), and reproduction (Inhibin level, luteal activity) (**Supplementary Table 12**).

Ongoing selection at functional loci

The Functional Annotation of Animal Genomes (FAANG) project has generated genome-wide data for multiple epigenetic marks (CTCF, ATAC, H3K4me1, H3K4me3, H3K27me3, and H3K27ac), in multiple tissues (adipose, cerebellum, cortex, hypothalamus, liver, lung, muscle and spleen) in a pair of biological replicates [280]. These data, coupled with sequence-level GPSM results allowed us to predict how selection has acted on functional regions of the genome. Previous selection mapping studies could generally resolve significant loci to positional candidate genes, but GPSM combined with COJO allowed us to refine the selected variant to the single base pair driving the signal. Eighty-seven COJO SNPs (35%) from the full Red Angus dataset resided in at least one of these epigenetic classes in at least one tissue. In the young Red Angus dataset, 166 of the COJO SNPs resided in one of these regions (40%). Certain GPSM SNPs clearly represent selection on regulatory variation. For example, a significant COJO SNP on chromosome 26 at 49,432,811 bp is immediately downstream of the gene Glutaredoxin-3 (*GLRX3*). The region containing this SNP harbors all six types of epigenetic marks, and these marks appear in all eight tissue types. Regulation of *GLRX3* has been postulated to play roles in feed efficiency [53] and calving ease [281]. While the regulatory site near *GLRX3* is highly diverse, most other regions are more specialized in their mark type and tissue context.

To determine if selected loci were more likely to play functional roles in the genome, we annotated COJO SNPs with their Functional and Evolutionary Trait Heritability (FAETH) scores as described in in Xiang et al. (2019) [282]. SNPs in the top ⅓ of the FAETH score distribution (per-variant score > 1.607×10^{-8}) were considered likely to be functional, as in Xiang et al. In Red Angus, 48% and 45% of COJO SNPs had FAETH scores in the top ⅓ of the FAETH score distribution for the full and young Red Angus datasets, respectively. A similar pattern existed in both Simmental datasets, where

46% and 39% of FAETH-annotated COJO SNPs were likely functional in the purebred and full dataset, respectively. This suggests that COJO SNPs identified by GPSM are more likely to be functional than SNPs chosen at random (t-test $p = 0.017$).

GPSM identifies breed-specific balancing selection in KHDRBS2 regulatory regions

The most significant GPSM locus in both Red Angus datasets resided between 1 and 2 Mb on Chromosome 23 (**Figure 4A-B**). The locus contains multiple sub-associations with birth date. In the young dataset, SNPs throughout the proximal 2 Mb of chromosome 23 are genome-wide significant. In this locus, 117 SNPs had p-values $< 10^{-310}$, reported by GCTA as zero. This included the most significant SNP in the full dataset (23:1,768,070, $p = 3.86 \times 10^{-94}$). This SNP was also the most significant COJO SNP in the young Red Angus dataset, suggesting that it is the variant responsible for the signal at this locus in both Red Angus datasets. Interestingly, it was not the most significant COJO SNP in the full Simmental dataset. Rather a SNP ~70 kb away (23:1,032,665, $p = 1.70 \times 10^{-247}$) had the strongest association (**Figure 4C**).

Forty-one SNPs within this locus were also significant in the full Simmental dataset, including 6 independent COJO associations. The lead Simmental SNP at this locus resides in a secondary association identified in the young Red Angus dataset. As in Red Angus, the most significant COJO SNP (23:1215338:G:T, COJO $p = 4.98 \times 10^{-238}$, raw $p = 4.28 \times 10^{-30}$) was not the most significant raw GPSM p-value (23:1364693:A:T, raw $p = 9.66 \times 10^{-38}$). Additionally, the strongest Simmental association for this locus (between 1.2 and 1.3 Mb) differed from the strongest association identified in Red Angus

(~1.7 Mb) (**Figure 4D**). This region does, however, overlap with one of the significant sub-associations found in Red Angus.

The closest protein coding gene to this signature is *KHDRBS2*, a gene involved in reproduction in goats [283], and calving ease in cattle [284]. *KHDRBS2* has also been identified as possessing a selection signature that differs between *Bos taurus* and *Bos indicus* cattle [285,286]. While our most significant COJO SNPs do not reside directly within known epigenetic mark regions (**Figure 4E**), dozens of marks exist within this 1 Mb segment and contain multiple genome-wide significant SNPs, suggesting strong ongoing selection for the regulation of *KHDRBS2* expression. Further, the annotated and expressed pseudogene *LOC782044* is 24,991 base pairs from the most significant GPSM association in Red Angus at 1,768,070 base pairs and 38,847 base pairs from a significant COJO SNP in Simmental. While *LOC782044* does not have any known functions in cattle, we hypothesize that it may be an enhancer RNA, aimed at altering the expression of *KHDRBS2* or other nearby genes [287].

The observed allele frequencies at significant COJO SNPs in this locus show two major patterns. First, we observe small directional changes in frequency, consistent with the polygenic selection that we observe at other GPSM loci (**Figure 4 F**). Second, for some Red Angus COJO SNPs, we observe allele frequencies oscillating around an intermediate value, a pattern consistent with balancing selection [288].

Sweep mapping identifies known and novel selected loci

GPSM and traditional selective sweep methods are both focused on identifying allele frequency changes and/or the signatures that they leave on surrounding neutral sites. We used two selection mapping methods, a haplotype-based statistic, number of segregating loci (nSL), and a composite statistic (μ) implemented by the software RAiSD. For nSL, we defined windows using the GenWin R package [289] which uses a spline function to observe changes in test statistics, and called the top 0.5% of windows significant, provided they contained at least three SNPs. Our nSL analysis in Red Angus identified 365 significant windows on all but three chromosomes (17,27, and 28) (**Supplementary Table 13, Supplementary Figure 6**). The correlation between nSL scores and GPSM effect sizes for 811K SNPs was -0.007. None of the significant GPSM COJO SNPs resided in outlier nSL windows. These significant nSL windows contained 134 annotated genes, 78 of which were identified by the selection signature meta-assembly in Randhawa et al. (2016). We used RAiSD to look for site frequency spectrum differences indicative of selection. RAiSD calculates μ in 50 SNP sliding windows (mean length = 39,139 bp), and we consider windows with the top 0.5% of μ values significant (**Supplementary Table 14**). The RAiSD analysis of sequenced Red Angus animals in the Thousand Bulls Project (n = 14) [290] identified 3,740 significant windows, many of which were overlapping, that encompassed dozens of loci exhibiting sweep-like signatures.

We found that 50% of significant nSL windows and 53% of significant RAiSD windows contained annotated genes (**Table 2**), a considerably higher proportion than identified by GPSM (between 30% and 34% across datasets). Of the 3,740 significant windows identified by RAiSD, 17 showed overlap with three distinct COJO associations in the young Red Angus dataset, but we did not observe any overlap with COJO SNPs in the full Red Angus dataset. We expect that In Red Angus, there were at least 12 sweep regions identified by both RAiSD and nSL. These included a locus on chromosome 6 (~78.95 Mb) that is a pleiotropic QTL for stayability, calving ease, and udder structure in dairy cattle, traits all under selection in the Red Angus breed [284]. Another shared sweep region on chromosome 11 (18.1 Mb) is associated with multiple carcass quality traits [236].

While the associated regions were largely different from those identified by GPSM analysis, we identified twelve GPSM COJO SNPs (4 in full, 8 in young Red Angus datasets, respectively) that reside within 50 kb of a significant nSL window. Two of these loci reside on chromosome 14 (23.0 Mb and 58.1 Mb). The locus at 23.0 Mb was also identified as a significant sweep region by RAiSD. This locus resides within the gene *TMEM68* which has previously been identified as a driver of feed intake and growth phenotypes in cattle [291], and height in human [174]. This locus also resides within a QTL for Insulin-like growth factor 1 level [292]. The locus at 58.1 Mb lies within the gene Oxidation Resistance 1 (*OXR1*), which is also known regulator of carcass weight in cattle [293] and neutrophil counts in humans [294]. A final shared region on chromosome 12 includes a region located immediately downstream of the gene *DNAJC15*, a heat shock protein with multiple reproductive associations in cattle [295,296] and human birth weight [297].

We expect that selective sweeps assert different pressures on neighboring neutral sites compared with polygenic selection. To quantify these differences, we calculated linkage disequilibrium r^2 statistics for all imputed SNPs within 100 kb of the lead SNP in significant RAiSD (n = 35) and nSL (n = 40) windows and significant GPSM COJO SNPs (n = 49) on chromosome 2 (**Figure 2E**). On average the r^2 in regions surrounding sweep loci (nSL loci mean $r^2 = 0.155$ [sd = 0.205], RAiSD loci mean $r^2 = 0.223$ [sd = 0.324]) were significantly higher (Tukey HSD p-value $< 2 \times 10^{-16}$) than those around GPSM loci (mean $r^2 = 0.105$ [sd = 0.206]).

Similarly, low levels of overlap between GPSM and sweep-detection methods existed in Simmental. A single GPSM COJO SNP resided within a significant RAiSD window (chromosome 26, 637,478 base pairs). The lone overlap between nSL and GPSM COJO SNPs in Simmental was on chromosome 5 at 57,701,350 base pairs. This is approximately 500 kb from the *PMEL*/*ERBB3* coat color candidate locus, and resides within a cluster of olfactory-associated genes. RAiSD and nSL analyses in Simmental identified at least 14 shared regions of selection (significant windows < 50 kb apart). This complementary evidence our confidence that an actual sweep occurred at a locus. The most notable shared locus between nSL and RAiSD is located at the *POLLED* locus [43] on chromosome 1, responsible for the presence or absence of horns. This locus has been under strong selection in most cattle populations, and is frequently identified in selection mapping studies [23,264]. Another shared region of selection on chromosome 16 at ~42.6 Mb near the genes *MASP2* and *TARDBP* has been identified in numerous other sweep mapping studies of Simmental cattle [18,176,219,220].

While overlapping nSL and RAiSD signatures can help bolster confidence in candidate selective sweeps. Most windows are uniquely identified by a single method (**Supplementary Tables 13-16**). For example, the most significant locus identified by RAiSD (**Supplementary Figure 7**) is located between two major clusters of T cell receptors on chromosome 10 at ~24.2 Mb. This sweep region was also identified in European Simmental populations by Qanbari et al. (2014) and Zhao et al. (2015)

[18,176], illustrating the long-lasting signatures created by some selective sweeps. The next most significant RAiSD windows in Simmental are located immediately upstream of *MC1R*, the gene responsible for red versus black coat color in cattle [298].

Ongoing polygenic selection acts on networks and pathways to drive complex trait changes

Using annotated positional candidate genes from each of our analyses, we performed protein-protein interaction network analysis to understand how selected candidate genes interacted with one another in a network. In both Red Angus datasets, networks had significantly more "edges" (protein-protein interactions, co-expression, co-localizations, and text-mining hits) than expected by chance (full Red Angus p = 0.034 , young Red Angus p = 5.84 x 10^{-6}). Networks were similarly enriched in the full (protein-protein interaction p = 0.00399) and purebred (protein-protein interaction p = 0.00903) Simmental datasets. In all cases, these networks had high average node degrees (the number of interactions per protein in the network), ranging from 1.05 (full Red Angus) to 2.34 (full Simmental) These imply relatively strong biological connections between candidate genes near loci actively under selection. STRING networks built with genes in significant nSL windows were not significant in Red Angus (p = 0.205), and marginally significant in Simmental (p = 0.027), though with significantly lower average node degrees (Red Angus = 0.985, Simmental = 0.962) than the GPSM networks. Networks of RAiSD candidate genes were the most significant (Red Angus p = 1.2 x 10^{-11}, Simmental p = 2.25 x 10^{-6}) and had high average node degrees (Red Angus = 2.37, Simmental = 3.35).

Discussion

In this study we further demonstrate the power of linear mixed models applied to a novel dependent variable to detect ongoing selection in populations with temporal genotype data. The Generation Proxy Selection Mapping (GPSM) method [13,14] is unique among selection mapping methods because it does not rely on outlier definitions, and significance is calculated on a per-marker basis, allowing us to pinpoint selection to the base pair level of resolution. Building on the work of Rowan et al. (2020), we expand GPSM to significantly larger datasets (46,454 and 78,787 animals) with imputed sequence level variants (> 11 million SNPs) [14]. This boost in power and resolution allowed us to map hundreds of small directional shifts in allele frequency, consistent with polygenic selection [299]. Further, by using a genomic relationship matrix, we are better able to control for the extensive population and family structure that exist in populations.

Due to the non-normality of generation proxy phenotypes in genotyped livestock populations, we explored the effects of transformations and data subsetting on GPSM's ability to detect selection. Large residuals for individuals born in the distant past led to a reduced power to detect selection in the full Red Angus dataset. When subsetting this dataset to individuals born very recently, or log-transforming the "birth date" generation proxy, we detected three times as many birth date-associated SNPs. In populations with low effective sizes (N_e), we might expect that stochastic changes in allele frequency due to drift could generate detectable changes in frequency [300], but simulations performed in our previous work have shown that GPSM is effectively able to distinguish between drift and selection [14]. While a log -transformed birth date generation proxy boosted the significance of signals in Red Angus, it did not have the same effect when applied to Simmental animals. As a result, for sequence-level analyses we used untransformed birth date as our dependent variable but partitioned the data into subsets to probe different components of selection.

Identifying selection at the base pair-level with conditional-joint analysis (COJO) allowed us to overlay functional data generated by the Functional Annotation of Animal Genomes (FAANG) project onto the independently associated GPSM SNPs. We map hundreds of selected loci with predicted epigenetic marks in cattle [280]. This selection on SNPs within epigenetic marks and other likely cis-regulatory regions suggests that selection on complex phenotypes is altering gene expression in complex networks [301–303]. Further, protein-protein interaction analyses showed significantly enriched biological connectedness, both for GPSM candidate genes and genes residing with nSL and RAiSD signatures.

While most other tests of polygenic selection explicitly test for correlations between allele frequencies and phenotypes over sampled time periods [304–306], or between distinct diverged populations [215,306], GPSM operates agnostic to phenotype and population label. This expands the odds of detecting polygenic selection signatures but can make interpreting signals difficult. Fortunately, hundreds of QTL-mapping studies have been performed in cattle, providing an extensive database of loci associated with economically-important complex traits [307]. We queried the Animal QTL Database for QTL near GPSM associations and performed enrichment tests to understand the traits under selection in these populations. In Red Angus, we identified significant QTL enrichments for several production traits such as body weight, average daily gain, and carcass weight, all traits that we would expect to find under selection in beef populations. We also identified multiple QTL classes influencing maternal traits and calving ease, two major recent selection emphases in the Red Angus breed. By analyzing both the full Simmental dataset and a subset of purebred animals, we could disentangle which allele frequencies were changing due to Simmental-specific selection versus selection on variation introgressed from other breeds. In purebred Simmental, we found limited selection on traits that were not explicitly involved in appearance characteristics. Strong selection at the *PMEL*/*ERBB3* and *KIT* loci have been a major focus of the breed as it aimed at making animals appear more like Angus cattle. As a result, less selection on complex production traits was detected by GPSM in the purebred dataset compared to the full dataset where we detected an excess of associations with carcass, production, and reproductive traits, consistent with ongoing selection in the cattle industry at large. Differences in the enriched carcass and production traits are consistent with traits that show appreciable average phenotypic differences between Angus and Simmental animals. While these QTL databases are far from comprehensive due to their biases towards loci with detectable effect sizes in frequently-measured traits, they provide a valuable first step to identifying the production traits driving genomic changes in these populations.

While some overlap between selective sweep mapping outlier windows (nSL and RAiSD) and GPSM hits existed, they are largely identifying different genomic loci. Sweep mapping methods consistently identify important Mendelian loci, such as *POLLED*, coat color genes (*MC1R*, *KIT*, etc.), and large-effect QTL where selection has for all intents and purposes been completed. GPSM's strength is in detecting subtle, directional shifts in allele frequency over short periods of time where selection is on-going. In contrast, nSL, RAiSD, and other sweep-mapping methods identify characteristics of sweeps such as extended haplotype homozygosity, long-range LD, and changes to the local site frequency spectrum. We observed significantly less LD between GPSM SNPs and neighboring sites compared to the lead SNPs in sweep peaks identified by nSL and RAiSD. In general, most sweep mapping strategies search for signatures at neighboring neutral sites, whereas GPSM tracks actual allele frequency changes over time.

Conclusions

Using large, commercially-generated cattle genotype datasets imputed to 11 million SNPs and the Generation Proxy Selection Mapping method we mapped hundreds of loci undergoing subtle directional shifts in frequency at the base pair level of resolution. These reside overwhelmingly in, or nearby, genes, which suggests that selection on complex traits is likely concerned with perturbing gene expression patterns in complex networks. As expected, GPSM detected largely different sets of selected loci than the selective sweep mapping methods. When longitudinally-sampled genotypes are available, GPSM is a powerful method for detecting ongoing changes to the genome. This makes it a complementary approach to sweep-mapping strategies as we work to understand the impacts of all types of selection on the genome.

Methods

Genotype data and imputation

We used commercially-generated assay genotypes from two populations of *Bos taurus* beef cattle. These data, made up of assays ranging in density from 25K to 777K SNPs were filtered, phased, then imputed using the approach described in Rowan et al. (2019) [7]. Briefly, prior to phasing and imputation we removed individuals with low call rates (< 0.90) and SNPs with low call rates (< 0.90) or extreme Hardy-Weinberg p-values ($< 10^{-50}$, indicative of genotyping errors) in PLINK (version 1.9) [61]. Genotypes were phased using a high-density reference in Eagle v2.4.1 [64] and imputed using Minimac4 [66]. The resulting high-density chip-imputed dataset contained 811,967 autosomal SNPs for 90,580 Simmental and 46,454 Red Angus animals. We refer to this dataset as 811K throughout. High-density imputed genotypes were then imputed to 43,214,290 SNPs from whole-genome resequencing data using 4,931 reference individuals from the Thousand Bulls Project Run8 [290]. We restricted the imputation reference to high quality (Variant Quality Score Recalibration [308] Tranche 90), biallelic variants with minor allele counts greater than 20. Following imputation, we further filtered imputed SNPs based on internally-calculated imputation R^2 (> 0.4) and minor allele frequency (> 0.01), leaving 11,759,568 imputed sequence variants for downstream analysis. Genomic coordinates for all array and sequence genotypes were based on positions in the ARS-UCD1.2 assembly [62].

Generation Proxy Phenotypes

We use the breeder-reported birth date to calculate birth date as the number of days of each animal's birth from the first-born animal within each dataset (as of October

19, 2020) for each animal to use as a continuous generation proxy in GPSM. In addition, due to the extreme left-skewness of animal birth dates in the dataset, we performed square root, cube root, and log transformations on animal birth date to test the effects of transforming the data for normality.

Generation Proxy Selection Mapping (GPSM)

Generation proxy selection mapping uses an individual's generation number, or a proxy for generation number as the dependent variable in a genomic variance component analysis or a genome-wide linear mixed model. To control for shared ancestry between individuals and to estimate variance components we used autosomal SNP markers in our 811K imputed dataset with MAF > 0.01 to construct a genomic relationship matrix (GRM) with the method in Yang et al. (2011) [121] for each population. Variance components were estimated using a genomic restricted maximum-likelihood (GREML) approach implemented in GCTA (version 1.92.3) [121,268]. To evaluate the impact of transformations to generation proxy phenotypes, we predicted random genetic effects (breeding values) and residuals for individuals using GCTA's "--reml-pred-rand" function.

The model used for selection mapping was as follows:

$$y = \mu + bx + a + \epsilon$$

Here, y is a vector of animal generation numbers or generation proxies, μ is the sample mean, bx is a the vector of regression coefficients b on an M x N matrix of animal genotypes x, a is a random vector of polygenic terms $\sim N(0, G\sigma_g{}^2)$ where G is a genomic relationship matrix, and $\epsilon$is a random error term $\sim N(0, I\sigma_e{}^2)$. All GPSM analyses were performed using the "--mlma" function in GCTA. When testing the impact of generation proxy transformations on statistical power, we used 811K imputed genotypes. We performed GPSM on sequence-imputed genotypes on four total datasets with GRMs calculated with 811K genotypes. The four datasets were as follows: the full Red Angus dataset (n = 46,454), Red Angus animals born on or after January 1st, 2012 (n = 44,470), all Simmental animals with at least 5% pedigree-reported Simmental ancestry (n = 78,787) and all purebred Simmental animals (n = 13,379).

To further refine GPSM signals and detect additional associations, we performed a within-analysis conditional and joint analysis (COJO) [269] in GCTA (v 1.92.3). COJO utilized summary data and genotypes from each of our sequence-level GPSM runs. The COJO model was conditioned on SNPs with GPSM p-values < 10^{-5}. We controlled for SNP collinearity by setting conditional p-values of highly-correlated variants ($r^2 > 0.9$) to 1.

Haplotype-based scans for selection

To map genomic regions that underwent strong selection in the distant-to-intermediate past, we used the number of segregating loci (nSL) method [309] on phased haplotypes from our full Red Angus and Simmental 811K datasets (MAF > 0.01), as well as on a subset of purebred Simmental animals. The nSL statistic was implemented in selscan [310] where per-SNP scores were also normalized in 100 frequency bins. As opposed to calculating significance in fixed windows, we fit a smoothing spline for each chromosome over normalized nSL scores using the GenWin R package [289]. This allowed us to define variable-length windows in which we calculated the mean nSL scores. We considered the top 0.5% of these windows to be significant outliers for downstream annotation and analysis.

Site frequency spectrum (SFS) scan for selection

We performed SFS-based scans for selection using called sequence genotypes from 14 American Red Angus and 32 registered Simmental animals, some of which may be crossbred, in the 1,000 Bull Genomes Project [290] in RAiSD (v2.9) [311]. RAiSD calculates a composite selection statistic, μ, aimed at detecting various signatures of selective sweeps. To preserve the full site frequency spectrum, we did not filter on any quality or frequency-based metrics. Rather, we restricted our analysis to only biallelic SNPs. SNPs in the top 0.5% of RAiSD μ values were considered significant outliers.

Gene & QTL annotation and enrichment

We annotated positional candidate genes and QTL using the GALLO R package with gene lists from ENSEMBL(Version 101) [312] and known *Bos taurus* QTL curated in the Animal QTL Database [307]. We annotated all genes and QTL within 50 kb of significant ($p < 5 \times 10^{-8}$) COJO SNPs or sweep outlier regions (based on lead SNP). We performed QTL enrichment analysis with GALLO, using an FDR-corrected p-value threshold of 0.05 to identify traits whose known QTL were significantly enriched in GPSM signatures.. We also annotated these SNPs with their Functional-And-Evolutionary Trait Heritability (FAETH) score as calculated in [282], with coordinates lifted over from the UMD3.1.1 assembly to ARS-UCD1.2 using the UCSC LiftOver tool [313].

FIGURES

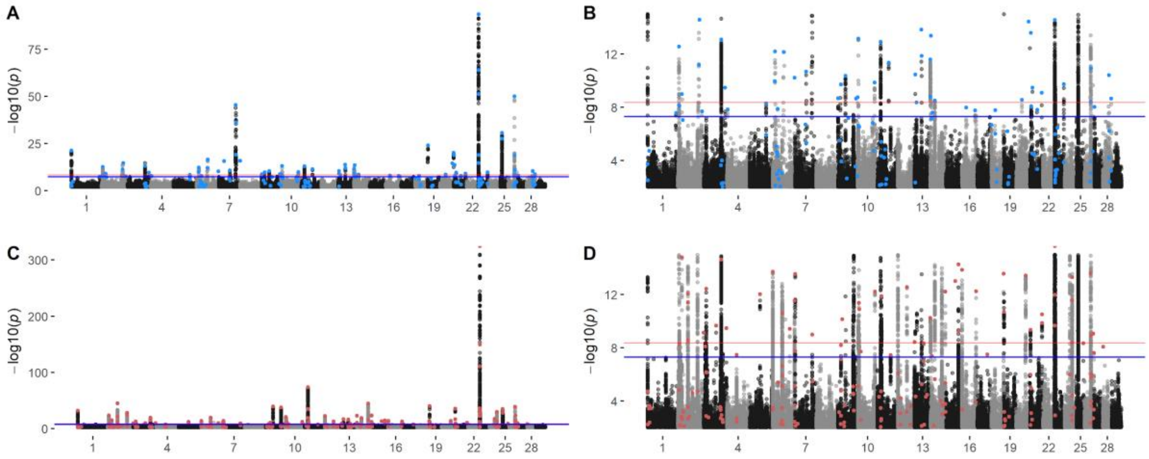


Figure 4.1. Red Angus GPSM Manhattan plots for the (A) full dataset, (B) truncated at p = 10^{-5} and (C) young dataset also truncated (D) at p = 10^{-5}. Genome-wide significance is indicated by the red line (Bonferroni-corrected p-value = 4.29 x 10^{-9}) and blue line (p-value = 5 x 10^{-8}). Blue points are significant COJO SNPs (COJO p < 5 x 10^{-5}) in the full Red Angus dataset. Red points are significant COJO SNPs in young Red Angus datasets. [Figure 1 in text]

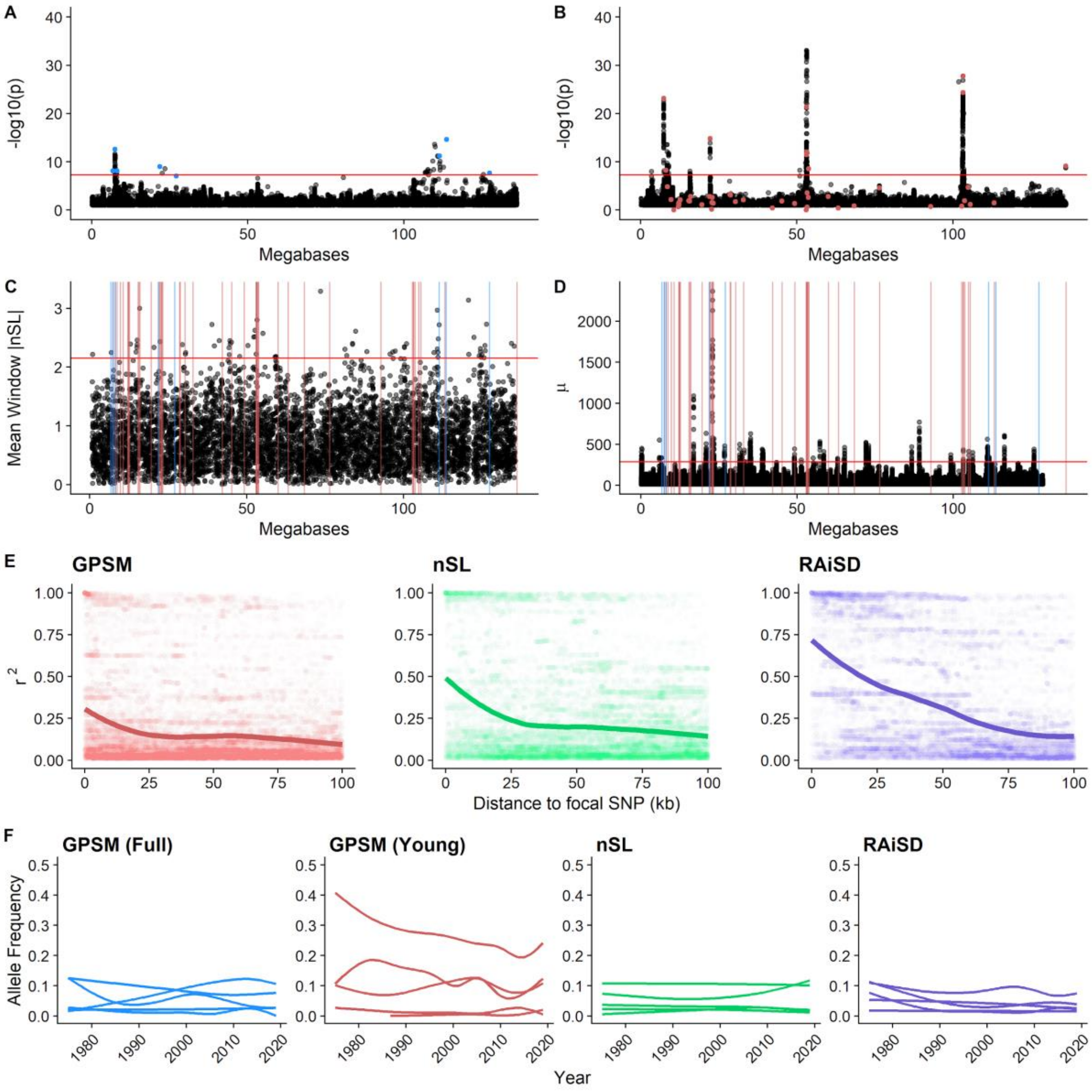


Figure 4.2. Methods identify largely different regions of selection on chromosome 2 in Red Angus. Chromosome 2 Manhattan plots for (A) full and (B) young Red Angus datasets. Red line is a Bonferroni-corrected significance threshold (4.29×10^{-9}). (C) Genomic distribution of |nSL|scores for windows defined by GenWin and (D) RAiSD μ statistics. Horizontal red lines indicate 0.5% outlier thresholds. Vertical red and blue lines represent the positions of full Red Angus and young Red Angus GPSM COJO SNPs, respectively. (E) Pairwise linkage disequilibrium (r^2) values for SNPs within 100 kb of

COJO SNPs for GPSM, SNPs closest to the center of significant nSL windows, or lead SNPs in significant RAiSD peaks. (F) Allele frequency trajectories over time for the five most significant SNPs in each analysis.

[Figure 2 in text]

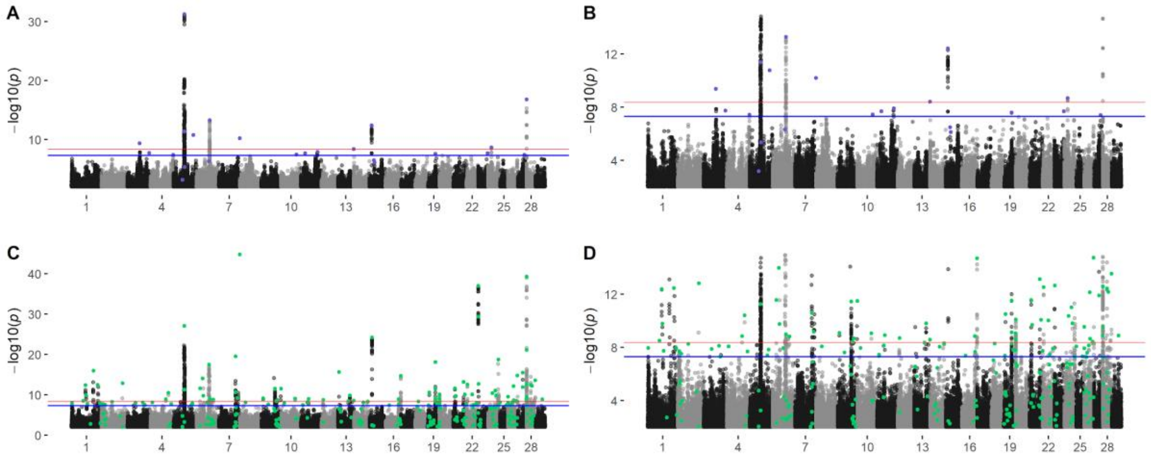


Figure 4.3. GPSM Manhattan plots for the (A) purebred Simmental dataset, (B) truncated at $p = 10^{-5}$ and (C) full Simmental dataset also truncated (D) at $p = 10^{-5}$. Genome-wide significance is indicated by the red line (Bonferroni-corrected p-value = 4.29×10^{-9}) and blue line (p-value = 5×10^{-8}). Purple points indicate significant COJO SNPs (COJO $p < 5 \times 10^{-5}$) in the purebred Simmental dataset. Green points are significant COJO SNPs in the full Simmental dataset.

[Figure 3 in text]

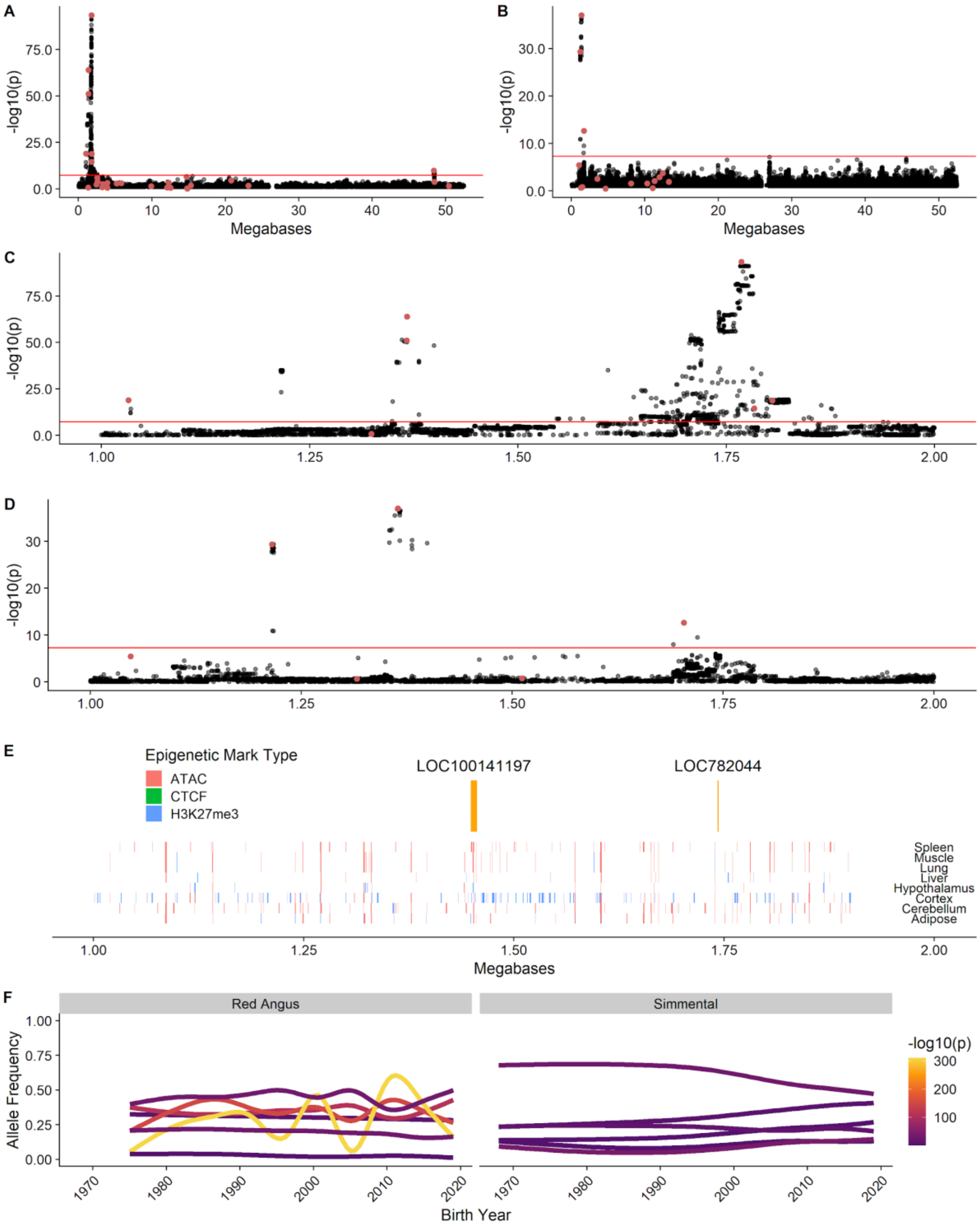


Figure 4.4. Manhattan plots for chromosome 23 in full (A) Red Angus and (B) Simmental datasets. Focused Manhattan plots at significant locus from 1-2 Mb on Chromosome 23 in full (C) Red Angus and (D) Simmental datasets. Red SNPs are

significant GPSM COJO associations (COJO $p < 5 \times 10^{-8}$). In A-D, the red line indicates a Bonferroni-corrected significance threshold. (E) Chromosome 23 (1-2 Mb) annotated with genes (orange) and epigenetic marks in eight tissues from two bovine samples in the FAANG project, colored by mark type. (F) Allele frequency trajectories represented by smoothened regression lines of birth year versus allele frequency for significant COJO SNPs in this region in the young Red Angus and full Simmental datasets. Lines are colored by the SNP's $-\log_{10}(p)$ value from GPSM.

[Figure 4 in text]

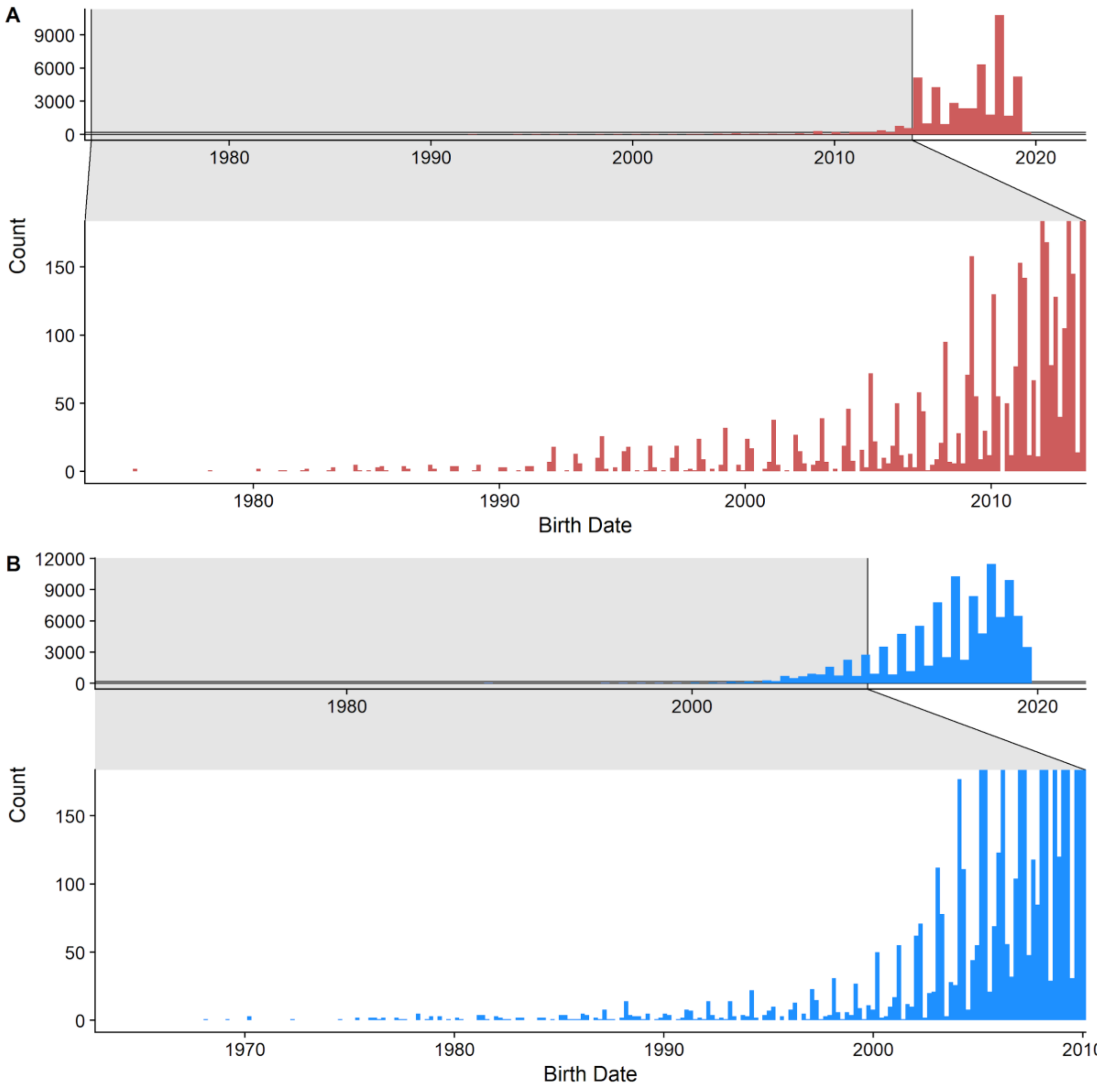


Figure 4.5. Distributions of birth dates in the full (A) Red Angus and (B) Simmental datasets. Faceted zoom on distant past individuals: 1975-2012 in Red Angus, 1965-2008 in Simmental.

[Supplementary Figure 1 in text]

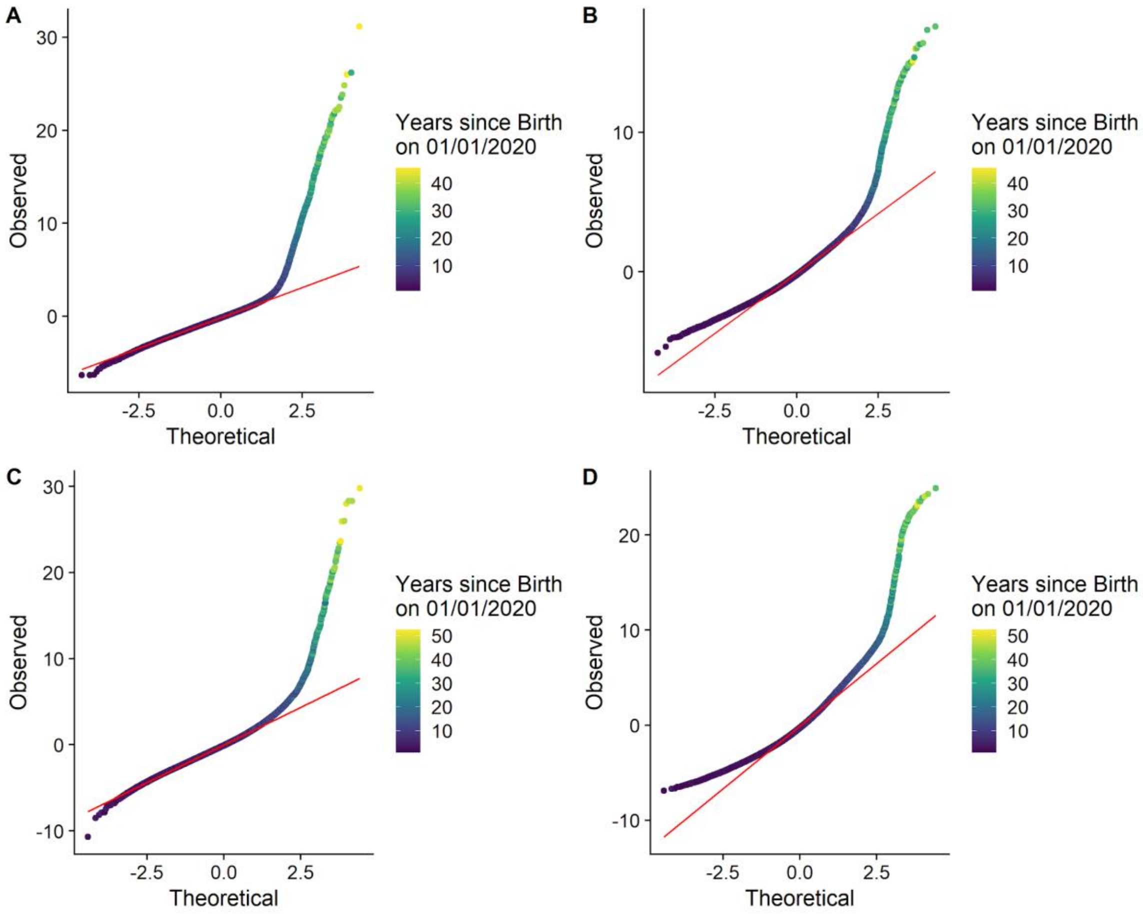


Figure 4.6. Residuals and breeding values for birth date GREML analysis. For Red Angus, estimated (A) residual error and (B) breeding values for each genotyped individual, colored by the number of years since birth. (C) Residual error and (D) breeding values for birth date in the full Simmental dataset.

[Supplementary Figure 2 in text]

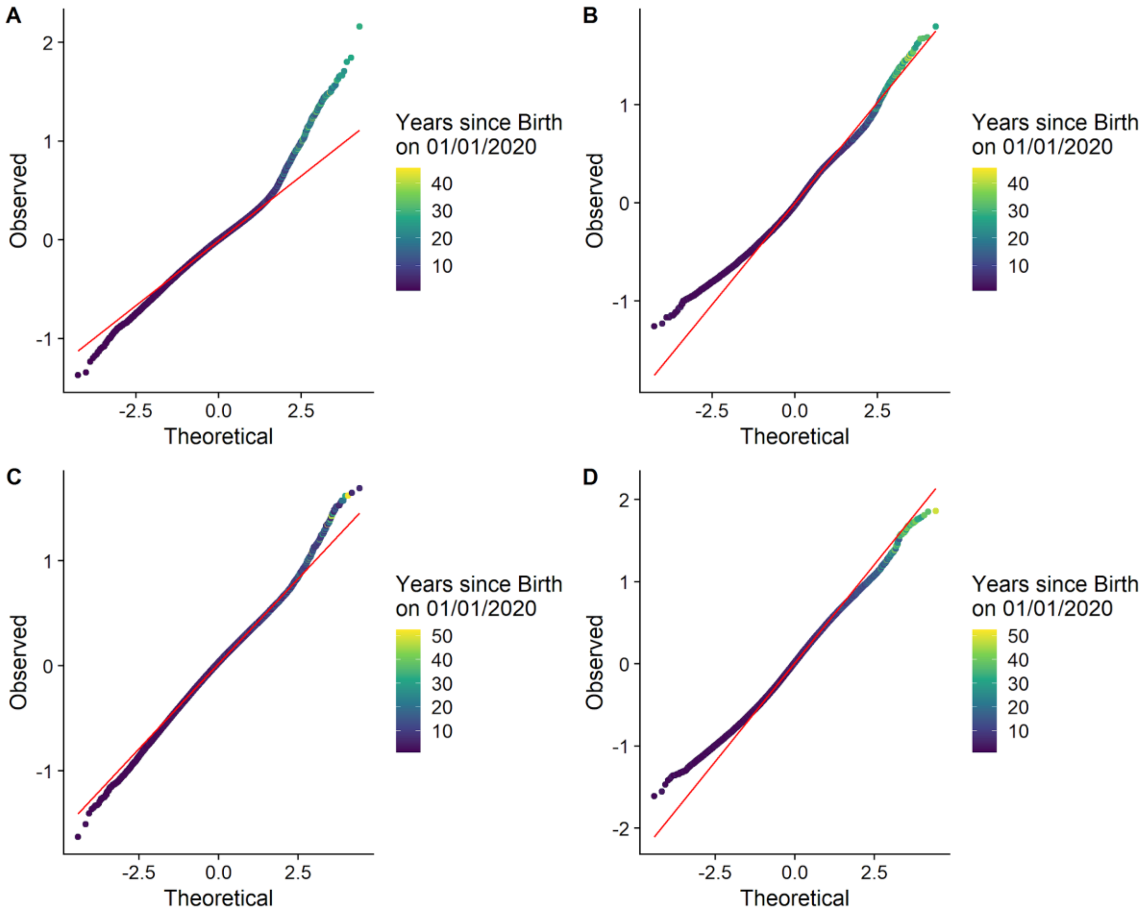


Figure 4.7. Residuals and breeding values for log-transformed years since birth date GREML analysis. For Red Angus, estimated (A) residual error and (B) breeding values for each genotyped individual, colored by the number of years since birth. (C) Residual error and (D) breeding values for birth date in the full Simmental dataset.

[Supplementary Figure 3 in text]

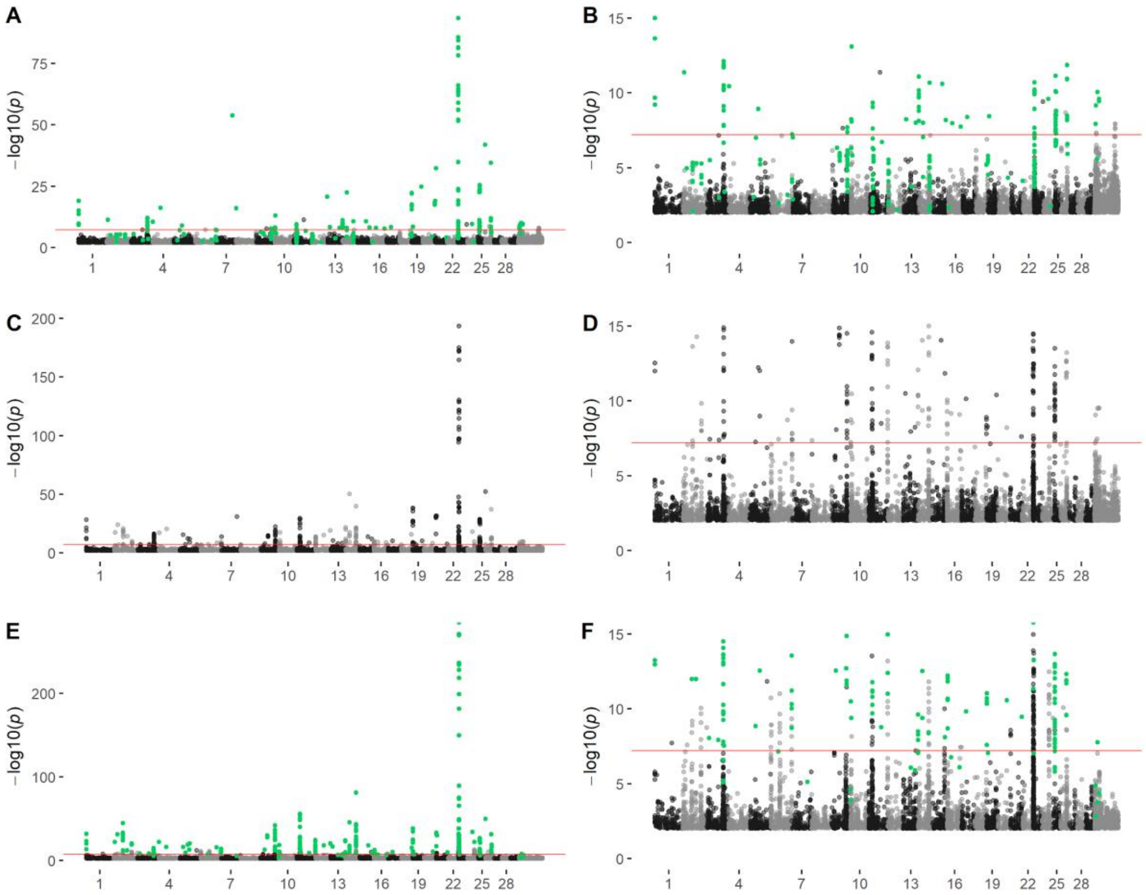


Figure 4.8. GPSM Manhattan plots comparing effects of transformation and data subsetting. 811K SNP Manhattan plots for Red Angus dataset using (A) raw birth date ((B) truncated at $-log_{10}(p) = 15$), (C) log-transformed birth date ((D) truncated at $-log_{10}(p) = 15$), or (E) raw birth dates for animals born since 2012 ((F) truncated at $-log_{10}(p) = 15$) as dependent variables in GPSM analysis. Genome-wide significance is indicated by the red line (Bonferroni-corrected p-value = 6.17 x 10^{-8}). Green points in A,B, E, and F represent genome-wide significant SNPs from the log-transformed age GPSM analysis. [Supplementary Figure 4 in text]

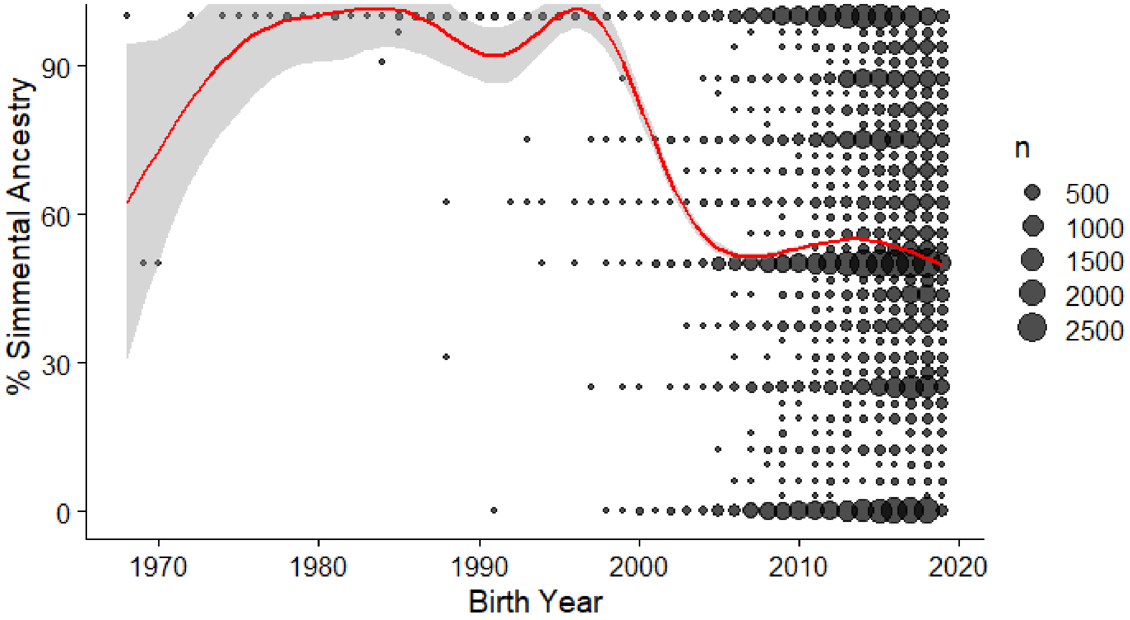


Figure 4.9. The changing breed composition of registered Simmental. Counts of Simmental ancestry percentages over time in all genotyped animals in American Simmental dataset. Points represent birth year/% Simmental ancestry combinations in the data, sized by the number of animals in each of those classes. The red line is a smoothed mean, surrounded by a 95% confidence interval in grey.

[Supplementary Figure 5 in text]

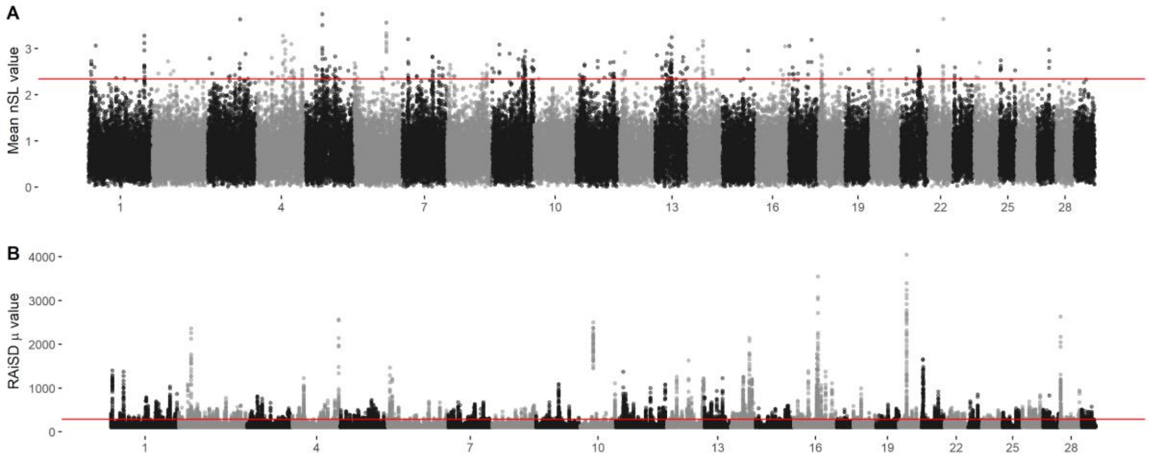


Figure 4.10. Selective sweep mapping in Red Angus. (A) Number of segregating loci (nSL) statistic windows across the genome. Window boundaries were defined by GenWin R package. Points represent the average |nSL| values within a window, with the genomic position defined as the genomic center of the window. Red line delineates 0.5% outliers deemed significant. (B) Manhattan plot of RAiSD μ statistics calculated from sequenced American Red Angus animals in the 1000 Bull Genomes Project. Red line delineates 0.5% outliers.

[Supplementary Figure 6 in text]

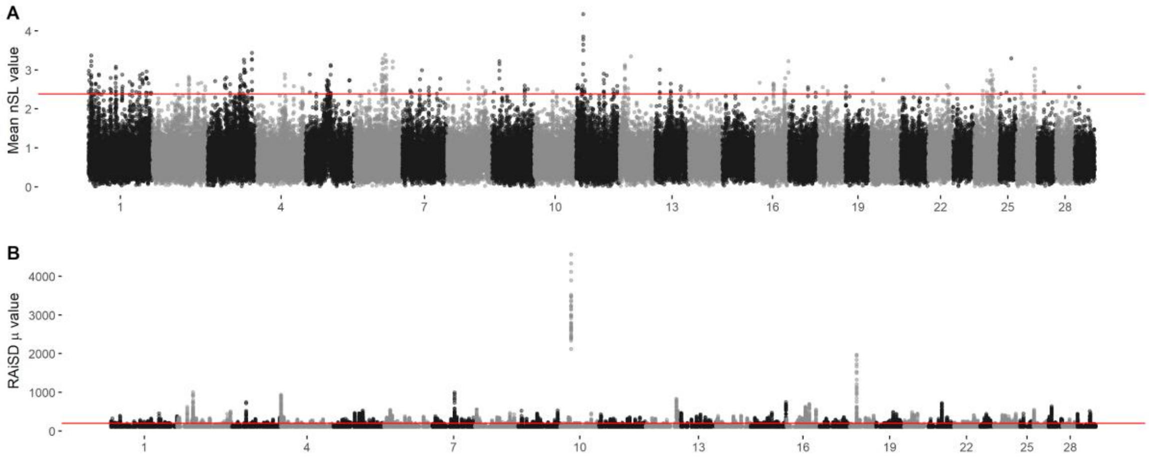


Figure 4.11. Selective sweep mapping in Simmental. (A) Number of segregating loci (nSL) statistic windows across the genome. Window boundaries were defined by GenWin R package. Points represent the average |nSL| values within a window, with the genomic position defined as the genomic center of the window. Red line delineates 0.5% outliers deemed significant. (B) Manhattan plot of RAiSD µ statistic calculated from sequenced American Simmental animals in the 1000 Bull Genomes Project. Red line delineates 0.5% outliers.

[Supplementary Figure 7 in text]

TABLES

Table 4.1. Number of significant ($p < 5 \times 10^{-8}$) COJO SNPs within, proximal, or outside of genic regions in each sequence-level dataset.

Breed	**Dataset**	**Total COJO SNPs**	**Within Gene**	**Gene Proximal (< 50kb to nearest)**	**Intergenic**
Red Angus	Full	248	76 (31%)	70 (28%)	102 (41%)
Red Angus	Young	417	142 (34%)	123 (29%)	152 (36%)
Simmental	Full	344	106 (30%)	119 (35%)	119 (35%)
Simmental	Purebred	33	10 (30%)	13 (40%)	10 (30%)

Table 4.2. Number of significant windows within, proximal, or outside of genic regions identified by nSL and RAiSD in each dataset.

Breed	**Analysis**	**Total Windows**	**Within Gene**	**Gene Proximal (< 50kb to nearest)**	**Intergenic**
Red Angus	nSL	365	183 (50%)	73 (20%)	109 (30%)
Red Angus	RAiSD	6,502	3,453 (53%)	811 (13%)	2,238 (34%)
Simmental	nSL	347	188 (54%)	60 (17%)	99 (29%)
Simmental	RAiSD	6,497	3,030 (47%)	1108 (17%)	2,359 (36%)

Table 4.3. Genomic restricted maximum likelihood (GREML) estimates of proportion variance explained (PVE) for various statistical transformations to birth date used as generation proxy in Red Angus.

Subset	**n(animals)**	**Dependent Variable**	**PVE (SE)**
FULL	46,454	Birth date	0.523 (0.007)
FULL	46,454	$\sqrt{birth\ date}$	0.616 (0.006)
FULL	46,454	$\sqrt[3]{birth\ date}$	0.637 (0.006)
FULL	46,454	$log(birth\ date)$	0.657 (0.006)
Young Only[1]	44,470	Birth date	0.773 (0.004)
Old Only[2]	1,984	Birth date	0.557 (0.031)

[1] Animals born on or after January 1, 2012
[2] Animals born prior to January 1, 2012

[Supplementary Table 1 in text]

Table 4.4. Genomic restricted maximum likelihood (GREML) estimates of proportion variance explained (PVE) for various subsets of the Simmental dataset.

Dataset	**n (animals)**	**Dependent Variable**	**PVE (Standard Error)**
Full[1]	78,787	Birth date	0.619 (0.005)
Full[1]	78,787	$log(birth\ date)$	0.600 (0.005)
Young Only[2]	73,811	Birth date	0.540 (0.005)
Old Only[3]	4,976	Birth date	0.436 (0.021)
<30% SIM, >50% AN[4]	11,429	Birth date	0.665 (0.011)
>20% SIM, <70%SIM[5]	46,136	Birth date	0.642 (0.006)
>70% SIM[6]	31,225	Birth date	0.558 (0.008)
>70% SIM[6]	31,225	$log(Birth\ date)$	0.561 (0.008)
Purebred[7]	13,379	Birth date	0.555 (0.011)
Purebred Young[8]	11,148,	Birth date	0.497 (0.013)
Purebred Old[9]	2,231	Birth date	0.462 (0.030)

[1] Animals with at least 5% Simmental ancestry
[2] Animals with at least 5% Simmental ancestry, born on or after January 1, 2008
[3] Animals with at least 5% Simmental ancestry, born prior to January 1, 2008
[4] Animals with less than 30% Simmental ancestry and more than 50% Angus ancestry
[5] Animals with more than 20%, but less than 70% Simmental ancestry
[6] Animals with more than 70% Simmental ancestry
[7] Animals with 100% Simmental ancestry
[8] Animals with 100% Simmental ancestry, born on or after January 1, 2008
[9] Animals with 100% Simmental ancestry, born prior to January 1, 2008

[Supplementary Table 2 in text]

Table 4.5. Counts of significant SNPs identified in each GPSM analysis of the Red Angus population using different transformations to generation proxy as dependent variable in 811K SNPs. The four significance cutoffs are reported: 1) A nominal significance ($p < 10^{-5}$), 2) a Bonferroni-adjusted threshold ($p < 7.55 \times 10^{-7}$), and FDR-corrected q-values 3) < 0.10 or 4) < 0.05.

Dataset	**n (animals)**	**Dependent Variable**	**nSNPs $p > 10^{-5}$ (nloci)**	**nSNPs $p < 7.55 \times 10^{-7}$ (nloci)**	**nSNPs $q < 0.10$ (nloci)**	**nSNPs $q < 0.05$ (nloci)**
Full	46,454	Birth date	315	214	509	398
Full	46,454	$\sqrt{Birth\ date}$	453	333	729	559
Full	46,454	$\sqrt[3]{Birth\ date}$	471	357	817	596
Full	46,454	$log(Birth\ date)$	513	377	822	715
Young Only[1]	44,470	Birth date	762	555	1210	1042
Old Only[2]	1,984	Birth date	18	1	1	1
Old Only[2]	1,984	$log(Birth\ date)$	3	0	0	0

[1] Animals born on or after January 1, 2012

[2] Animals born prior to January 1, 2012

[Supplementary Table 3 in text]

Table 4.6. Counts of significant SNPs identified in each GPSM analysis of the Simmental population using different population subsets and transformations to generation proxy with 811K SNPs. Ancestry proportions are pedigree estimates reported by American Simmental Association. The four significance thresholds are reported: 1) A nominal significance ($p < 10^{-5}$), 2) a Bonferroni-adjusted threshold ($p < 7.55 \times 10^{-7}$), and FDR-corrected q-values 3) < 0.10 or 4) < 0.05.

Dataset	**n (animals)**	**Dependent Variable**	**Nominal ($p < 10^{-5}$)**	**Bonferroni**	**q < 0.1**	**q < 0.05**
Full[1]	78,787	Birth date	120	70	137	117
Full[1]	78,787	$log(Birth\ date)$	109	77	130	105
Young Only[2]	73,811	Birth date	94	68	100	89
Old Only[3]	4,976	Birth date	119	72	171	115
<30% SIM, >50% AN[4]	11,429	Birth date	14	3	3	0
>20% SIM, <70%SIM[5]	46,136	Birth date	46	31	39	38
>70% SIM[6]	31,225	Birth date	100	61	107	88
>70% SIM[6]	31,225	$log(Birth\ date)$	51	24	44	32
Purebred[7]	13,379	Birth date	92	50	111	85
Purebred Young[8]	11,148,	Birth date	11	4	4	3
Purebred Old[9]	2,231	Birth date	60	34	59	48

[1] Animals with at least 5% Simmental ancestry
[2] Animals with at least 5% Simmental ancestry, born on or after January 1, 2008
[3] Animals with at least 5% Simmental ancestry, born prior to January 1, 2008
[4] Animals with less than 30% Simmental ancestry and more than 50% Angus ancestry
[5] Animals with more than 20%, but less than 70% Simmental ancestry
[6] Animals with more than 70% Simmental ancestry
[7] Animals with 100% Simmental ancestry
[8] Animals with 100% Simmental ancestry, born on or after January 1, 2008
[9] Animals with 100% Simmental ancestry, born prior to January 1, 2008

[Supplementary Table 4 in text]

Supplementary Tables 4.7-4.18 are located here:

https://docs.google.com/spreadsheets/d/1bxf9Cl3ZCsgkzycLu8MFYg6YyWPegwYVMntpqCydvJo/edit?usp=sharing

Table 4.7. Significant ($p < 5 \times 10^{-8}$) conditional and joint (COJO) SNPs from full Red Angus GPSM analysis and their annotated positional candidate genes (< 50 kb) [Supplementary Table 5 in text].

Table 4.8. Significant ($p < 5 \times 10^{-8}$) conditional and joint (COJO) SNPs from young Red Angus GPSM analysis and their annotated positional candidate genes (< 50 kb). [Supplementary Table 6 in text]

Table 4.9. Significantly enriched QTL in regions (< 50 kb) from significant GPSM COJO SNPs identified in the full Red Angus dataset (n SNPs = 248) . [Supplementary Table 7 in text]

Table 4.10. Significantly enriched QTL in regions (< 50 kb) from significant GPSM COJO SNPs identified in the young Red Angus dataset (n SNPs = 417). [Supplementary Table 8]

Table 4.11. Significant ($p < 5 \times 10^{-8}$) conditional and joint (COJO) SNPs from full Simmental GPSM analysis and their annotated positional candidate genes (< 50 kb). [Supplementary Table 9 in text]

Table 4.12. Significant ($p < 5 \times 10^{-8}$) conditional and joint (COJO) SNPs from purebred Simmental GPSM analysis and their annotated positional candidate genes (< 50 kb).

[Supplementary Table 10 in text]

Table 4.13. Significantly enriched QTL in regions (< 50 kb) from significant GPSM COJO SNPs identified in full Simmental dataset (n SNPs = 344) .

[Supplementary Table 11 in text]

Table 4.14. Significantly enriched QTL in regions (< 50 kb) from significant GPSM COJO SNPs identified in purebred Simmental dataset (n SNPs = 33) .

[Supplementary Table 12 in text]

Table 4.15. Red Angus outlier nSL windows (top 0.5%) as defined by GenWin R package. Genes that fell within significant windows are reported with their corresponding window.

[Supplementary Table 13 text]

Table 4.16. Unique genes within Red Angus outlier RAiSD windows (top 0.5%). Reported window is the window with the highest μ value that overlaps with the gene.

[Supplementary Table 14 in text]

Table 4.17. Red Angus outlier nSL windows (top 0.5%) as defined by GenWin R package. Genes that fell within significant windows are reported with their corresponding window.

[Supplementary Table 15]

Table 4.18. Unique genes within Simmental outlier RAiSD windows (top 0.05%). Reported window is the window with the highest μ value that overlaps with the gene.

[Supplementary Table 16 in text]

REFERENCES

1. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157: 1819–1829.

2. Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, et al. Development and characterization of a high density SNP genotyping assay for cattle. PLoS One. 2009;4: e5350.

3. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91: 4414–4423.

4. Taylor JF, Taylor KH, Decker JE. Holsteins are the genomic selection poster cows. Proceedings of the National Academy of Sciences of the United States of America. 2016. pp. 7690–7692.

5. García-Ruiz A, Cole JB, VanRaden PM, Wiggans GR, Ruiz-López FJ, Van Tassell CP. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proc Natl Acad Sci U S A. 2016;113: E3995–4004.

6. Genomic Evaluations. [cited 12 Nov 2020]. Available: https://www.uscdcb.com/what-we-do/genomics/

7. Rowan TN, Hoff JL, Crum TE, Taylor JF, Schnabel RD, Decker JE. A multi-breed reference panel and additional rare variants maximize imputation accuracy in cattle. Genet Sel Evol. 2019;51: 77.

8. Rolf MM, Taylor JF, Schnabel RD, McKay SD, McClure MC, Northcutt SL, et al. Impact of reduced marker set estimation of genomic relationship matrices on genomic selection for feed efficiency in Angus cattle. BMC Genet. 2010;11: 24.

9. Su G, Brøndum RF, Ma P, Guldbrandtsen B, Aamand GP, Lund MS. Comparison of genomic predictions using medium-density (~54,000) and high-density (~777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. J Dairy Sci. 2012;95: 4657–4665.

10. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010;11: 499–511.

11. MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile-Mariam M, Kemper KE, Chamberlain AJ, et al. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics. 2016;17: 144.

12. Köster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. Bioinformatics. 2012;28: 2520–2522.

13. Decker JE, Vasco DA, McKay SD, McClure MC, Rolf MM, Kim J, et al. A novel analytical method, Birth Date Selection Mapping, detects response of the Angus (Bos taurus) genome to selection on complex traits. BMC Genomics. 2012;13: 606.

14. Rowan TN, Durbin HJ, Seabury CM, Schnabel RD, Decker JE. Powerful detection of polygenic selection and environmental adaptation in US beef cattle. 2020. p. 2020.03.11.988121. doi:10.1101/2020.03.11.988121

15. Falconer DS. Introduction to quantitative genetics. Oliver And Boyd; Edinburgh; London; 1960.

16. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res. 1974;23: 23–35.

17. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. Localizing recent adaptive evolution in the human genome. PLoS Genet. 2007;3: e90.

18. Qanbari S, Pausch H, Jansen S, Somel M, Strom TM, Fries R, et al. Classic selective sweeps revealed by massive sequencing in cattle. PLoS Genet. 2014;10: e1004148.

19. Nair S, Williams JT, Brockman A, Paiphun L, Mayxay M, Newton PN, et al. A selective sweep driven by pyrimethamine treatment in southeast asian malaria parasites. Mol Biol Evol. 2003;20: 1526–1536.

20. Eyre-Walker A. Evolution in health and medicine Sackler colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proc Natl Acad Sci U S A. 2010;107 Suppl 1: 1752–1756.

21. Barghi N, Hermisson J, Schlötterer C. Polygenic adaptation: a unifying framework to understand positive selection. Nat Rev Genet. 2020. doi:10.1038/s41576-020-0250-z

22. Höllinger I, Pennings PS, Hermisson J. Polygenic adaptation: From sweeps to subtle frequency shifts. PLoS Genet. 2019;15: e1008035.

23. Xu L, Bickhart DM, Cole JB, Schroeder SG, Song J, Van Tassell CP, et al. Genomic signatures reveal new evidences for selection of important traits in domestic cattle. Mol Biol Evol. 2015;32: 711–725.

24. Randhawa IAS, Khatkar MS, Thomson PC, Raadsma HW. A Meta-Assembly of Selection Signatures in Cattle. PLoS One. 2016;11: e0153013.

25. Walsh B, Lynch M. Evolution and selection of quantitative traits. 2018. Available: https://books.google.ca/books?hl=en&lr=&id=L2liDwAAQBAJ&oi=fnd&pg=PP1&ots=y9dWVmdg1F&sig=pOREAZIAXXiV3gcMJ2WO-qKSEkc

26. St-Pierre NR, Cobanov B, Schnitkey G. Economic Losses from Heat Stress by US

Livestock Industries1. J Dairy Sci. 2003;86: E52–E77.

27. Ahlberg CM, Allwardt K, Broocks A, Bruno K, Taylor A, Mcphillips L, et al. Characterization of water intake and water efficiency in beef cattle1,2. J Anim Sci. 2019;97: 4770–4782.

28. Strickland JR, Aiken GE, Spiers DE, Fletcher LR, Oliver JW. Physiological Basis of Fescue Toxicosis. In: Fribourg HA, Hannaway DB, West CP, editors. Tall Fescue for the Twenty-first Century. Madison, WI, USA: American Society of Agronomy, Crop Science Society of America, Soil Science Society of America; 2009. pp. 203–227.

29. Bradford HL, Fragomeni BO, Bertrand JK, Lourenco DAL, Misztal I. Genetic evaluations for growth heat tolerance in Angus cattle. J Anim Sci. 2016;94: 4143–4150.

30. Carvalheiro R, Costilla R, Neves HHR, Albuquerque LG, Moore S, Hayes BJ. Unraveling genetic sensitivity of beef cattle to environmental variation under tropical conditions. Genet Sel Evol. 2019;51: 29.

31. Hayes BJ, Daetwyler HD, Goddard ME. Models for genome× environment interaction: Examples in livestock. Crop Sci. 2016;56: 2251–2259.

32. Smith JL, Wilson ML, Nilson SM, Rowan TN, Oldeschulte DL, Schnabel RD, et al. Genome-wide association and genotype by environment interactions for growth traits in U.S. Gelbvieh cattle. BMC Genomics. 2019;20: 926.

33. Braz CU, Rowan TN, Schnabel RD, Decker JE. Extensive genome-wide association analyses identify genotype-by-environment interactions of growth traits in Simmental cattle. doi:10.1101/2020.01.09.900902

34. Butts WT, Koger M, Pahnish OF, Burns WC, Warwick EJ. Performance of two lines of Hereford cattle in two environments. J Anim Sci. 1971;33: 923–932.

35. Savolainen O, Lascoux M, Merilä J. Ecological genomics of local adaptation. Nat Rev Genet. 2013;14: 807–820.

36. Günther T, Coop G. Robust identification of local adaptation from allele frequencies. Genetics. 2013;195: 205–220.

37. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: Genomic selection in dairy cattle: progress and challenges. J Dairy Sci. 2009;92: 433–443.

38. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, et al. Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci. 2009;92: 16–24.

39. Grisart B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P, et al. Positional

candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. 2002;12: 222–231.

40. Grisart B, Farnir F, Karim L, Cambisano N, Kim J-J, Kvasz A, et al. Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. Proc Natl Acad Sci U S A. 2004;101: 2398–2403.

41. Kambadur R, Sharma M, Smith TP, Bass JJ. Mutations in myostatin (GDF8) in double-muscled Belgian Blue and Piedmontese cattle. Genome Res. 1997;7: 910–916.

42. McPherron AC, Lee SJ. Double muscling in cattle due to mutations in the myostatin gene. Proc Natl Acad Sci U S A. 1997;94: 12457–12461.

43. Wiedemar N, Tetens J, Jagannathan V, Menoud A, Neuenschwander S, Bruggmann R, et al. Independent polled mutations leading to complex gene expression differences in cattle. PLoS One. 2014;9: e93435.

44. Ron M, Weller JI. From QTL to QTN identification in livestock--winning by points rather than knock-out: a review. Anim Genet. 2007;38: 429–439.

45. Saatchi M, Schnabel RD, Taylor JF, Garrick DJ. Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. BMC Genomics. 2014;15: 442.

46. Hoff JL, Decker JE, Schnabel RD, Taylor JF. Candidate lethal haplotypes and causal mutations in Angus cattle. BMC Genomics. 2017;18: 799.

47. Wiggans GR, Cooper TA, VanRaden PM, Van Tassell CP, Bickhart DM, Sonstegard TS. Increasing the number of single nucleotide polymorphisms used in genomic evaluation of dairy cattle. J Dairy Sci. 2016;99: 4504–4511.

48. Bouwman AC, Daetwyler HD, Chamberlain AJ, Ponce CH, Sargolzaei M, Schenkel FS, et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nat Genet. 2018;50: 362–367.

49. Wiggans GR, Cole JB, Hubbard SM, Sonstegard TS. Genomic Selection in Dairy Cattle: The USDA Experience. Annu Rev Anim Biosci. 2017;5: 309–327.

50. Decker JE. Agricultural Genomics: Commercial Applications Bring Increased Basic Research Power. PLoS Genet. 2015;11: e1005621.

51. Pausch H, MacLeod IM, Fries R, Emmerling R, Bowman PJ, Daetwyler HD, et al. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. Genet Sel Evol. 2017;49: 24.

52. Frischknecht M, Pausch H, Bapst B, Signer-Hasler H, Flury C, Garrick D, et al. Highly accurate sequence imputation enables precise QTL mapping in Brown Swiss cattle. BMC Genomics. 2017;18: 999.

53. Seabury CM, Oldeschulte DL, Saatchi M, Beever JE, Decker JE, Halley YA, et al. Genome-wide association study for feed efficiency and growth traits in U.S. beef cattle. BMC Genomics. 2017;18: 386.

54. Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012;95: 4114–4129.

55. Fang L, Sahana G, Ma P, Su G, Yu Y, Zhang S, et al. Use of biological priors enhances understanding of genetic architecture and genomic prediction of complex traits within and between dairy cattle breeds. BMC Genomics. 2017;18: 604.

56. Zhang Q, Sahana G, Su G, Guldbrandtsen B, Lund MS, Calus MPL. Impact of rare and low-frequency sequence variants on reliability of genomic prediction in dairy cattle. Genet Sel Evol. 2018;50: 62.

57. Whalen A, Gorjanc G, Ros-Freixedes R, Hickey JM. Assessment of the performance of hidden Markov models for imputation in animal breeding. Genet Sel Evol. 2018;50: 44.

58. van Binsbergen R, Bink MC, Calus MP, van Eeuwijk FA, Hayes BJ, Hulsegge I, et al. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol. 2014;46: 41.

59. Kreiner-Møller E, Medina-Gomez C, Uitterlinden AG, Rivadeneira F, Estrada K. Improving accuracy of rare variant imputation with a two-step imputation approach. Eur J Hum Genet. 2015;23: 395–400.

60. Pausch H, Emmerling R, Schwarzenbacher H, Fries R. A multi-trait meta-analysis with imputed sequence variants reveals twelve QTL for mammary gland morphology in Fleckvieh cattle. Genet Sel Evol. 2016;48: 14.

61. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81: 559–575.

62. Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience. 2020;9. doi:10.1093/gigascience/giaa021

63. Crum TE, Schnabel RD, Decker JE, Regitano LCA, Taylor JF. CRUMBLER: A tool for the Prediction of Ancestry in Cattle. bioRxiv. 2018. p. 396341. doi:10.1101/396341

64. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016;48: 1443–1448.

65. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27: 2987–2993.

66. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48: 1284–1287.

67. Lin P, Hartz SM, Zhang Z, Saccone SF, Wang J, Tischfield JA, et al. A new statistic to evaluate imputation reliability. PLoS One. 2010;5: e9697.

68. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44: 821–824.

69. Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014;46: 858–865.

70. Ramnarine S, Zhang J, Chen L-S, Culverhouse R, Duan W, Hancock DB, et al. When Does Choice of Accuracy Measure Alter Imputation Accuracy Assessments? PLoS One. 2015;10: e0137601.

71. Hancock DB, Levy JL, Gaddis NC, Bierut LJ, Saccone NL, Page GP, et al. Assessment of genotype imputation performance using 1000 Genomes in African American studies. PLoS One. 2012;7: e50610.

72. Hartl DL, Clark AG, Clark AG. Principles of population genetics. Sinauer associates Sunderland; 1997.

73. Druet T, Schrooten C, de Roos APW. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. J Dairy Sci. 2010;93: 5443–5454.

74. VanRaden PM, Null DJ, Sargolzaei M, Wiggans GR, Tooker ME, Cole JB, et al. Genomic imputation and evaluation using high-density Holstein genotypes. J Dairy Sci. 2013;96: 668–678.

75. Brøndum RF, Guldbrandtsen B, Sahana G, Lund MS, Su G. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. BMC Genomics. 2014;15: 728.

76. Bovine HapMap Consortium, Gibbs RA, Taylor JF, Van Tassell CP, Barendse W, Eversole KA, et al. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science. 2009;324: 528–532.

77. Gutiérrez-Gil B, Arranz JJ, Wiener P. An interpretive review of selective sweep

studies in Bos taurus cattle populations: identification of unique and shared selection signals across breeds. Front Genet. 2015;6: 167.

78. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, et al. Classic selective sweeps were rare in recent human evolution. Science. 2011;331: 920–924.

79. Ma Y, Ding X, Qanbari S, Weigend S, Zhang Q, Simianer H. Properties of different selection signature statistics and a new strategy for combining them. Heredity . 2015;115: 426–436.

80. Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, et al. Development and characterization of a high density SNP genotyping assay for cattle. PLoS One. 2009;4: e5350.

81. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet. 2014;15: 335–346.

82. Mészáros G, Petautschnig E, Schwarzenbacher H, Sölkner J. Genomic regions influencing coat color saturation and facial markings in Fleckvieh cattle. Anim Genet. 2015;46: 65–68.

83. Fontanesi L, Tazzoli M, Russo V, Beever J. Genetic heterogeneity at the bovine KIT gene in cattle breeds carrying different putative alleles at the spotting locus. Anim Genet. 2010;41: 295–303.

84. Kuehn LA, Thallman RM. Across-Breed EPD Tables For The Year 2016 Adjusted To Breed Differences For Birth Year Of 2014. 2016 [cited 9 Feb 2020]. Available: https://digitalcommons.unl.edu/hruskareports/380/

85. Weber KL, Welly BT, Van Eenennaam AL, Young AE, Porto-Neto LR, Reverter A, et al. Identification of Gene Networks for Residual Feed Intake in Angus Cattle Using Genomic Prediction and RNA-seq. PLoS One. 2016;11: e0152274.

86. Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, Phadnis N, et al. Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. PLoS Genet. 2009;5: e1000753.

87. Ma L, O'Connell JR, VanRaden PM, Shen B, Padhi A, Sun C, et al. Cattle Sex-Specific Recombination and Genetic Control from a Large Pedigree Analysis. PLoS Genet. 2015;11: e1005387.

88. Guo B, Greenwood PL, Cafe LM, Zhou G, Zhang W, Dalrymple BP. Transcriptome analysis of cattle muscle identifies potential markers for skeletal muscle growth rate and major cell types. BMC Genomics. 2015;16: 177.

89. Rolf MM, Taylor JF, Schnabel RD, McKay SD, McClure MC, Northcutt SL, et al. Genome-wide association analysis for feed efficiency in Angus cattle. Anim Genet.

2012;43: 367–374.

90. Breeds - Red Angus. In: The Cattle Site [Internet]. [cited 28 Feb 2020]. Available: https://www.thecattlesite.com/breeds/beef/99/red-angus/

91. Gareis NC, Huber E, Hein GJ, Rodríguez FM, Salvetti NR, Angeli E, et al. Impaired insulin signaling pathways affect ovarian steroidogenesis in cows with COD. Anim Reprod Sci. 2018;192: 298–312.

92. Davis SL, Hossner KL, Ohlson DL. Endocrine Regulation of Growth in Ruminants. In: Roche JF, O'Callaghan D, editors. Manipulation of Growth in Farm Animals: A Seminar in the CEC Programme of Coordination of Research on Beef Production, held in Brussels December 13–14, 1982. Dordrecht: Springer Netherlands; 1984. pp. 151–178.

93. Hill WG. Applications of population genetics to animal breeding, from wright, fisher and lush to genomic prediction. Genetics. 2014;196: 1–16.

94. Coop G, Witonsky D, Di Rienzo A, Pritchard JK. Using environmental correlations to identify loci underlying local adaptation. Genetics. 2010;185: 1411–1423.

95. Yoder JB, Stanton-Geddes J, Zhou P, Briskine R, Young ND, Tiffin P. Genomic signature of adaptation to climate in Medicago truncatula. Genetics. 2014;196: 1263–1275.

96. Li J, Chen G-B, Rasheed A, Li D, Sonder K, Zavala Espinosa C, et al. Identifying loci with breeding potential across temperate and tropical adaptation via EigenGWAS and EnvGWAS. Mol Ecol. 2019;28: 3544–3560.

97. Lasky JR, Upadhyaya HD, Ramu P, Deshpande S, Hash CT, Bonnette J, et al. Genome-environment associations in sorghum landraces predict adaptive traits. Sci Adv. 2015;1: e1400218.

98. Sathiaraj D, Huang X, Chen J. Predicting climate types for the Continental United States using unsupervised clustering techniques. Environmetrics. 2019. p. e2524. doi:10.1002/env.2524

99. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447: 661–678.

100. Sansregret L, Nepveu A. The multiple roles of CUX1: insights from mouse models and cell-based assays. Gene. 2008;412: 84–94.

101. Bertolini F, Servin B, Talenti A, Rochat E, Kim ES, Oget C, et al. Signatures of selection and environmental adaptation across the goat genome post-domestication. Genet Sel Evol. 2018;50: 57.

102. Aiken GE, Klotz JL, Looper ML, Tabler SF, Schrick FN. Disrupted hair follicle activity in cattle grazing endophyte-infected tall fescue in the summer insulates core body temperatures1. The Professional Animal Scientist. 2011;27: 336–343.

103. León CD, De León C, Manrique C, Martínez R, Rocha JF. Research Article Genomic association study for adaptability traits in four Colombian cattle breeds. Genetics and Molecular Research. 2019. doi:10.4238/gmr18373

104. Guo J, Tao H, Li P, Li L, Zhong T, Wang L, et al. Whole-genome sequencing reveals selection signatures associated with important traits in six goat breeds. Sci Rep. 2018;8: 10405.

105. Gurgul A, Jasielczuk I, Ropka-Molik K, Semik-Gurgul E, Pawlina-Tyszko K, Szmatoła T, et al. A genome-wide detection of selection signatures in conserved and commercial pig breeds maintained in Poland. BMC Genet. 2018;19: 95.

106. Medugorac I, Graf A, Grohs C, Rothammer S, Zagdsuren Y, Gladyr E, et al. Whole-genome analysis of introgressive hybridization and characterization of the bovine legacy of Mongolian yaks. Nat Genet. 2017;49: 470–475.

107. Guo D-F, Cui H, Zhang Q, Morgan DA, Thedens DR, Nishimura D, et al. The BBSome Controls Energy Homeostasis by Mediating the Transport of the Leptin Receptor to the Plasma Membrane. PLoS Genet. 2016;12: e1005890.

108. Davis RE, Swiderski RE, Rahmouni K, Nishimura DY, Mullins RF, Agassandian K, et al. A knockin mouse model of the Bardet–Biedl syndrome 1 M390R mutation has cilia defects, ventriculomegaly, retinopathy, and obesity. Proc Natl Acad Sci U S A. 2007;104: 19422–19427.

109. Ai H, Fang X, Yang B, Huang Z, Chen H, Mao L, et al. Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. Nat Genet. 2015;47: 217–225.

110. Morrison SF. Central control of body temperature. F1000Res. 2016;5. doi:10.12688/f1000research.7958.1

111. Garner JB, Douglas ML, Williams SRO, Wales WJ, Marett LC, Nguyen TTT, et al. Genomic Selection Improves Heat Tolerance in Dairy Cattle. Sci Rep. 2016;6: 34114.

112. Davies VJ, Hollins AJ, Piechota MJ, Yip W, Davies JR, White KE, et al. Opa1 deficiency in a mouse model of autosomal dominant optic atrophy impairs mitochondrial morphology, optic nerve structure and visual function. Hum Mol Genet. 2007;16: 1307–1318.

113. Patten DA, Wong J, Khacho M, Soubannier V, Mailloux RJ, Pilon-Larose K, et al. OPA1-dependent cristae modulation is essential for cellular adaptation to metabolic demand. EMBO J. 2014;33: 2676–2691.

114. Gopalakrishnan K, Kumarasamy S, Abdul-Majeed S, Kalinoski AL, Morgan EE, Gohara AF, et al. Targeted disruption of Adamts16 gene in a rat genetic model of hypertension. Proc Natl Acad Sci U S A. 2012;109: 20555–20559.

115. Dong K, Yao N, Pu Y, He X, Zhao Q, Luan Y, et al. Genomic scan reveals loci under altitude adaptation in Tibetan and Dahe pigs. PLoS One. 2014;9: e110520.

116. Fraser HB. Gene expression drives local adaptation in humans. Genome Res. 2013;23: 1089–1096.

117. Castric V, Bechsgaard J, Schierup MH, Vekemans X. Repeated adaptive introgression at a gene under multiallelic balancing selection. PLoS Genet. 2008;4: e1000168.

118. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Methods. 2014;11: 407–409.

119. R Core Team R, Others. R: A language and environment for statistical computing. R foundation for statistical computing Vienna, Austria; 2013.

120. Wickham H. ggplot2. Wiley Interdisciplinary Reviews: Computational. 2011. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.147

121. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88: 76–82.

122. PRISM Climate Group. PRISM 30-year Normal Climate Data. Available: http://prism.oregonstate.edu

123. Dudbridge F, Gusnanto A. Estimation of significance thresholds for genomewide association scans. Genet Epidemiol. 2008;32: 227–234.

124. Han B, Eskin E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. Am J Hum Genet. 2011;88: 586–598.

125. Han B, Eskin E. Interpreting meta-analyses of genome-wide association studies. PLoS Genet. 2012;8: e1002555.

126. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009;25: 1091–1093.

127. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13: 2498–2504.

128. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et

al. Proteomics. Tissue-based map of the human proteome. Science. 2015;347: 1260419.

129. Mouse ENCODE Consortium, Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, et al. An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol. 2012;13: 418.

130. Jain A, Tuteja G. TissueEnrich: Tissue-specific gene enrichment analysis. Bioinformatics. 2019;35: 1966–1967.

131. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015;348: 648–660.

132. Xu X, Wells AB, O’Brien DR, Nehorai A, Dougherty JD. Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders. J Neurosci. 2014;34: 1420–1431.

133. Kim W, Underwood RS, Greenwald I, Shaye DD. OrthoList 2: A New Comparative Genomic Analysis of Human and Caenorhabditis elegans Genes. Genetics. 2018;210: 445–461.

134. Angeles-Albores D, N Lee RY, Chan J, Sternberg PW. Tissue enrichment analysis for C. elegans genomics. BMC Bioinformatics. 2016;17: 366.

135. Angeles-Albores D, Lee RYN, Chan J, Sternberg PW. Two new functions in the WormBase Enrichment Suite. microPublication Biology: https://doi. org/10.17912. W17925Q17912N; 2018.

136. Noyes H, Brass A, Obara I, Anderson S, Archibald AL, Bradley DG, et al. Genetic and expression analysis of cattle identifies candidate genes in pathways responding to Trypanosoma congolense infection. Proc Natl Acad Sci U S A. 2011;108: 9304–9309.

137. Smetko A, Soudre A, Silbermayr K, Müller S, Brem G, Hanotte O, et al. Trypanosomosis: potential driver of selection in African cattle. Front Genet. 2015;6: 137.

138. Álvarez I, Pérez-Pardal L, Traoré A, Fernández I, Goyache F. African Cattle do not Carry Unique Mutations on the Exon 9 of the ARHGAP15 Gene. Anim Biotechnol. 2016;27: 9–12.

139. Serão NV, González-Peña D, Beever JE, Faulkner DB, Southey BR, Rodriguez-Zas SL. Single nucleotide polymorphisms and haplotypes associated with feed efficiency in beef cattle. BMC Genet. 2013;14: 94.

140. Müller M-P, -P. Müller M, Rothammer S, Seichter D, Russ I, Hinrichs D, et al. Genome-wide mapping of 10 calving and fertility traits in Holstein dairy cattle with special regard to chromosome 18. Journal of Dairy Science. 2017. pp. 1987–2006.

doi:10.3168/jds.2016-11506

141. Seo J, Osorio JS, Loor JJ. Purinergic signaling gene network expression in bovine polymorphonuclear neutrophils during the peripartal period. J Dairy Sci. 2013;96: 7675–7683.

142. Chelh I, Picard B, Hocquette J-F, Cassar-Malek I. Myostatin inactivation induces a similar muscle molecular signature in double-muscled cattle as in mice. Animal. 2011;5: 278–286.

143. Seo J, Osorio JS, Schmitt E, Corrêa MN, Bertoni G, Trevisi E, et al. Hepatic purinergic signaling gene network expression and its relationship with inflammation and oxidative stress biomarkers in blood from peripartal dairy cattle. J Dairy Sci. 2014;97: 861–873.

144. Espigolan R, Baldi F, Boligon AA, Souza FRP, Fernandes Júnior GA, Gordo DGM, et al. Associations between single nucleotide polymorphisms and carcass traits in Nellore cattle using high-density panels. Genet Mol Res. 2015;14: 11133–11144.

145. Derks MFL, Gjuvsland AB, Bosse M, Lopes MS, van Son M, Harlizius B, et al. Loss of function mutations in essential genes cause embryonic lethality in pigs. PLoS Genet. 2019;15: e1008055.

146. Meade KG, Gormley E, O’Farrelly C, Park SD, Costello E, Keane J, et al. Antigen stimulation of peripheral blood mononuclear cells from Mycobacterium bovis infected cattle yields evidence for a novel gene expression program. BMC Genomics. 2008;9: 447.

147. Cassar-Malek I, Boby C, Picard B, Reverter A, Hudson NJ. Molecular regulation of high muscle mass in developing Blonde d’Aquitaine cattle foetuses. Biol Open. 2017;6: 1483–1492.

148. Höglund JK, Guldbrandtsen B, Lund MS, Sahana G. Analyzes of genome-wide association follow-up study for calving traits in dairy cattle. BMC Genet. 2012;13: 71.

149. Chen Y, Gondro C, Quinn K, Herd RM, Parnell PF, Vanselow B. Global gene expression profiling reveals genes expressed differentially in cattle with high and low residual feed intake. Anim Genet. 2011;42: 475–490.

150. Marete AG, Guldbrandtsen B, Lund MS, Fritz S, Sahana G, Boichard D. A Meta-Analysis Including Pre-selected Sequence Variants Associated With Seven Traits in Three French Dairy Cattle Populations. Front Genet. 2018;9: 522.

151. Evangelou E, Warren HR, Mosen-Ansorena D, Mifsud B, Pazoki R, Gao H, et al. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. Nat Genet. 2018;50: 1412–1425.

152. Hussin J, Nadeau P, Lefebvre J-F, Labuda D. Haplotype allelic classes for detecting ongoing positive selection. BMC Bioinformatics. 2010;11: 65.

153. Wang G, Yang E, Smith KJ, Zeng Y, Ji G, Connon R, et al. Gene expression responses of threespine stickleback to salinity: implications for salt-sensitive hypertension. Front Genet. 2014;5: 312.

154. Barbaux S, Gascoin-Lachambre G, Buffat C, Monnier P, Mondon F, Tonanny M-B, et al. A genome-wide approach reveals novel imprinted genes expressed in the human placenta. Epigenetics. 2012;7: 1079–1090.

155. Boitard S, Boussaha M, Capitan A, Rocha D, Servin B. Uncovering Adaptation from Sequence Data: Lessons from Genome Resequencing of Four Cattle Breeds. Genetics. 2016;203: 433–450.

156. Gautier M, Moazami-Goudarzi K, Levéziel H, Parinello H, Grohs C, Rialle S, et al. Deciphering the Wisent Demographic and Adaptive Histories from Individual Whole-Genome Sequences. Mol Biol Evol. 2016;33: 2801–2814.

157. Andrade WA, Silva AM, Alves VS, Salgado APC, Melo MB, Andrade HM, et al. Early endosome localization and activity of RasGEF1b, a toll-like receptor-inducible Ras guanine-nucleotide exchange factor. Genes Immun. 2010;11: 447–457.

158. Lopez M, Choin J, Sikora M, Siddle K, Harmant C, Costa HA, et al. Genomic Evidence for Local Adaptation of Hunter-Gatherers to the African Rainforest. Curr Biol. 2019;29: 2926–2935.e4.

159. Wang MD, Dzama K, Hefer CA, Muchadeyi FC. Genomic population structure and prevalence of copy number variations in South African Nguni cattle. BMC Genomics. 2015;16: 894.

160. Alkorta-Aranburu G, Beall CM, Witonsky DB, Gebremedhin A, Pritchard JK, Di Rienzo A. The genetic architecture of adaptations to high altitude in Ethiopia. PLoS Genet. 2012;8: e1003110.

161. Delgado-Benito V, Rosen DB, Wang Q, Gazumyan A, Pai JA, Oliveira TY, et al. The Chromatin Reader ZMYND8 Regulates Igh Enhancers to Promote Immunoglobulin Class Switch Recombination. Mol Cell. 2018;72: 636–649.e8.

162. Gong F, Chiu L-Y, Cox B, Aymard F, Clouaire T, Leung JW, et al. Screen identifies bromodomain protein ZMYND8 in chromatin recognition of transcription-associated DNA damage that promotes homologous recombination. Genes Dev. 2015;29: 197–211.

163. Hirota T, Takahashi A, Kubo M, Tsunoda T, Tomita K, Sakashita M, et al. Genome-wide association study identifies eight new susceptibility loci for atopic dermatitis in the Japanese population. Nat Genet. 2012;44: 1222–1226.

164. Srikanth K, Kwon A, Lee E, Chung H. Characterization of genes and pathways that respond to heat stress in Holstein calves through transcriptome analysis. Cell Stress Chaperones. 2017;22: 29–42.

165. Rahman MB, Kamal MM, Rijsselaere T, Vandaele L, Shamsuddin M, Van Soom A. Altered chromatin condensation of heat-stressed spermatozoa perturbs the dynamics of DNA methylation reprogramming in the paternal genome after in vitro fertilisation in cattle. Reprod Fertil Dev. 2014;26: 1107–1116.

166. Aramaki M, Kimura T, Udaka T, Kosaki R, Mitsuhashi T, Okada Y, et al. Embryonic expression profile of chicken CHD7, the ortholog of the causative gene for CHARGE syndrome. Birth Defects Res A Clin Mol Teratol. 2007;79: 50–57.

167. Nakajima T, Wooding S, Satta Y, Jinnai N, Goto S, Hayasaka I, et al. Evidence for natural selection in the HAVCR1 gene: high degree of amino-acid variability in the mucin domain of human HAVCR1 protein. Genes Immun. 2005;6: 398–406.

168. McLoughlin KE, Nalpas NC, Rue-Albrecht K, Browne JA, Magee DA, Killick KE, et al. RNA-seq Transcriptional Profiling of Peripheral Blood Leukocytes from Cattle Infected with Mycobacterium bovis. Front Immunol. 2014;5: 396.

169. Kim J, Williams FJ, Dreger DL, Plassais J, Davis BW, Parker HG, et al. Genetic selection of athletic success in sport-hunting dogs. Proc Natl Acad Sci U S A. 2018;115: E7212–E7221.

170. Rodríguez A, Rusciano T, Hamilton R, Holmes L, Jordan D, Wollenberg Valero KC. Genomic and phenotypic signatures of climate adaptation in an Anolis lizard. Ecol Evol. 2017;7: 6390–6403.

171. Wang H, Ding K, Zhang Y, Jin L, Kullo IJ, He F. Comparative and evolutionary pharmacogenetics of ABCB1: complex signatures of positive selection on coding and regulatory regions. Pharmacogenet Genomics. 2007;17: 667–678.

172. Wang Z, Wang J, Tantoso E, Wang B, Tai AYP, Ooi LLPJ, et al. Signatures of recent positive selection at the ATP-binding cassette drug transporter superfamily gene loci. Hum Mol Genet. 2007;16: 1367–1380.

173. López Herráez D, Bauchet M, Tang K, Theunert C, Pugach I, Li J, et al. Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. PLoS One. 2009;4: e7888.

174. Kichaev G, Bhatia G, Loh P-R, Gazal S, Burch K, Freund MK, et al. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. Am J Hum Genet. 2019;104: 65–75.

175. Tachmazidou I, Süveges D, Min JL, Ritchie GRS, Steinberg J, Walter K, et al. Whole-Genome Sequencing Coupled to Imputation Discovers Genetic Signals for Anthropometric Traits. Am J Hum Genet. 2017;100: 865–884.

176. Zhao F, McParland S, Kearney F, Du L, Berry DP. Detection of selection signatures in dairy and beef cattle using high-density genomic information. Genet Sel Evol. 2015;47: 49.

177. Cardoso DF, de Albuquerque LG, Reimer C, Qanbari S, Erbe M, do Nascimento AV, et al. Genome-wide scan reveals population stratification and footprints of recent selection in Nelore cattle. Genet Sel Evol. 2018;50: 22.

178. Rubin C-J, Zody MC, Eriksson J, Meadows JRS, Sherwood E, Webster MT, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature. 2010;464: 587–591.

179. Stone S, Abkevich V, Russell DL, Riley R, Timms K, Tran T, et al. TBC1D1 is a candidate for a severe obesity gene and evidence for a gene/gene interaction in obesity predisposition. Hum Mol Genet. 2006;15: 2709–2720.

180. Chadt A, Leicht K, Deshmukh A, Jiang LQ, Scherneck S, Bernhardt U, et al. Tbc1d1 mutation in lean mouse strain confers leanness and protects from diet-induced obesity. Nat Genet. 2008;40: 1354–1359.

181. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell. 2016;167: 1415–1429.e19.

182. Northrup JM, Shafer ABA, Anderson CR, Coltman DW, Wittemyer G. Fine-scale genetic correlates to condition and migration in a wild cervid. Evol Appl. 2014;7: 937–948.

183. Duforet-Frebourg N, Bazin E, Blum MGB. Genome scans for detecting footprints of local adaptation using a Bayesian factor model. Mol Biol Evol. 2014;31: 2483–2495.

184. Igoshin AV, Yurchenko AA, Belonogova NM, Petrovsky DV, Aitnazarov RB, Soloshenko VA, et al. Genome-wide association study and scan for signatures of selection point to candidate genes for body temperature maintenance under the cold stress in Siberian cattle populations. BMC Genet. 2019;20: 26.

185. Wang B, Zhang Y-B, Zhang F, Lin H, Wang X, Wan N, et al. On the origin of Tibetans and their genetic basis in adapting high-altitude environments. PLoS One. 2011;6: e17002.

186. Amorim CEG, Daub JT, Salzano FM, Foll M, Excoffier L. Detection of convergent genome-wide signals of adaptation to tropical forests in humans. PLoS One. 2015;10: e0121557.

187. Granka JM, Henn BM, Gignoux CR, Kidd JM, Bustamante CD, Feldman MW. Limited evidence for classic selective sweeps in African populations. Genetics. 2012;192: 1049–1064.

188. Flori L, Moazami-Goudarzi K, Alary V, Araba A, Boujenane I, Boushaba N, et al. A genomic map of climate adaptation in Mediterranean cattle breeds. Mol Ecol. 2019;28: 1009–1029.

189. Shin S-Y, Fauman EB, Petersen A-K, Krumsiek J, Santos R, Huang J, et al. An atlas of genetic influences on human blood metabolites. Nat Genet. 2014;46: 543–550.

190. Key FM, Fu Q, Romagné F, Lachmann M, Andrés AM. Human adaptation and population differentiation in the light of ancient genomes. Nat Commun. 2016;7: 10775.

191. Mueller JC, Pulido F, Kempenaers B. Identification of a gene associated with avian migratory behaviour. Proc Biol Sci. 2011;278: 2848–2856.

192. Bazzi G, Galimberti A, Hays QR, Bruni I, Cecere JG, Gianfranceschi L, et al. Adcyap1 polymorphism covaries with breeding latitude in a Nearctic migratory songbird, the Wilson’s warbler (Cardellina pusilla). Ecol Evol. 2016;6: 3226–3239.

193. Jones SE, Lane JM, Wood AR, van Hees VT, Tyrrell J, Beaumont RN, et al. Genome-wide association analyses of chronotype in 697,828 individuals provides insights into circadian rhythms. Nat Commun. 2019;10: 343.

194. Sjöstrand AE, Sjödin P, Jakobsson M. Private haplotypes can reveal local adaptation. BMC Genet. 2014;15: 61.

195. Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ. Genomic patterns of homozygosity in worldwide human populations. Am J Hum Genet. 2012;91: 275–292.

196. Rouabhi M, Guo DF, Rahmouni K. Bardet-Biedl Syndrome 1 Gene in the Ventromedial Hypothalamus is Required for Energy Homeostasis. The FASEB Journal. 2016;30: 750.4–750.4.

197. Dickinson RE, Duncan WC. The SLIT-ROBO pathway: a regulator of cell function with implications for the reproductive system. Reproduction. 2010;139: 697–704.

198. Wu D-D, Zhang Y-P. Positive selection drives population differentiation in the skeletal genes in modern humans. Hum Mol Genet. 2010;19: 2341–2346.

199. Chen Z. Physiological, transcriptomic and genomic mechanisms of thermal adaptation in Oncorhynchus mykiss. University of British Columbia. 2017. doi:10.14288/1.0340726

200. do Prado FD, Vera M, Hermida M, Bouza C, Pardo BG, Vilas R, et al. Parallel evolution and adaptation to environmental factors in a marine flatfish: Implications for fisheries and aquaculture management of the turbot (Scophthalmus maximus).

Evol Appl. 2018;11: 1322–1341.

201. Twomey AJ, Berry DP, Evans RD, Doherty ML, Graham DA, Purfield DC. Genome-wide association study of endo-parasite phenotypes using imputed whole-genome sequence data in dairy and beef cattle. Genet Sel Evol. 2019;51: 15.

202. Li RW, Li C. Butyrate induces profound changes in gene expression related to multiple signal pathways in bovine kidney epithelial cells. BMC Genomics. 2006;7: 234.

203. Ringseis R, Zeitz JO, Weber A, Koch C, Eder K. Hepatic transcript profiling in early-lactation dairy cows fed rumen-protected niacin during the transition from late pregnancy to lactation. J Dairy Sci. 2019;102: 365–376.

204. Williams JL, Dunner S, Valentini A, Mazza R, Amarger V, Checa ML, et al. Discovery, characterization and validation of single nucleotide polymorphisms within 206 bovine genes that may be considered as candidate genes for beef production and quality. Anim Genet. 2009;40: 486–491.

205. Weldenegodguad M, Popov R, Pokharel K, Ammosov I, Ming Y, Ivanova Z, et al. Whole-Genome Sequencing of Three Native Cattle Breeds Originating From the Northernmost Cattle Farming Regions. Front Genet. 2018;9: 728.

206. Lei N, Mellem JE, Brockie PJ, Madsen DM, Maricq AV. NRAP-1 Is a Presynaptically Released NMDA Receptor Auxiliary Protein that Modifies Synaptic Strength. Neuron. 2017;96: 1303–1316.e6.

207. Wang Y, Harashima S-I, Liu Y, Usui R, Inagaki N. Sphingosine kinase 1-interacting protein is a novel regulator of glucose-stimulated insulin secretion. Sci Rep. 2017;7: 779.

208. Cañadas-Garre M, Anderson K, Cappa R, Skelly R, Smyth LJ, McKnight AJ, et al. Genetic Susceptibility to Chronic Kidney Disease - Some More Pieces for the Heritability Puzzle. Front Genet. 2019;10: 453.

209. Purfield DC, Bradley DG, Evans RD, Kearney FJ, Berry DP. Genome-wide association study for calving performance using high-density genotypes in dairy and beef cattle. Genet Sel Evol. 2015;47: 47.

210. Ang CE, Ma Q, Wapinski OL, Fan S, Flynn RA, Lee QY, et al. The novel lncRNA lnc-NR2F1 is pro-neurogenic and mutated in human neurodevelopmental disorders. Elife. 2019;8. doi:10.7554/eLife.41770

211. Lu W-C, Zhou Y-X, Qiao P, Zheng J, Wu Q, Shen Q. The protocadherin alpha cluster is required for axon extension and myelination in the developing central nervous system. Neural Regeneration Res. 2018;13: 427–433.

212. Yang R, Fang S, Wang J, Zhang C, Zhang R, Liu D, et al. Genome-wide analysis

of structural variants reveals genetic differences in Chinese pigs. PLoS One. 2017;12: e0186721.

213. Wright S. THE GENETICAL STRUCTURE OF POPULATIONS. Ann Eugen. 1949;15: 323–354.

214. Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, Blott S, et al. Detecting selection in population trees: the Lewontin and Krakauer test extended. Genetics. 2010;186: 241–262.

215. Chen H, Patterson N, Reich D. Population differentiation as a test for selective sweeps. Genome Res. 2010;20: 393–402.

216. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007;449: 913–918.

217. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002;419: 832–837.

218. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006;4: e72.

219. Ramey HR, Decker JE, McKay SD, Rolf MM, Schnabel RD, Taylor JF. Detection of selective sweeps in cattle using genome-wide SNP data. BMC Genomics. 2013;14: 382.

220. Rothammer S, Seichter D, Förster M, Medugorac I. A genome-wide scan for signatures of differential artificial selection in ten cattle breeds. BMC Genomics. 2013;14: 908.

221. Utsunomiya YT, Pérez O’Brien AM, Sonstegard TS, Van Tassell CP, do Carmo AS, Mészáros G, et al. Detecting loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods. PLoS One. 2013;8: e64280.

222. Drouillard JS. Current situation and future trends for beef production in the United States of America — A review. Asian-Australasian Journal of Animal Sciences. 2018. pp. 1007–1016. doi:10.5713/ajas.18.0428

223. Burns WC, Koger M, Butts WT, Pahnish OF, Blackwell RL. Genotype by Environment Interaction in Hereford Cattle: II. Birth and Weaning Traits. J Anim Sci. 1979;49: 403–409.

224. Hohenboken W, Jenkins T, Pollak J, Bullock D, Radakovich S. Genetic improvement of beef cattle adaptation in America. Proceedings of the Beef Improvement Federation’s 37th annual research symposium and annual meeting.

2005. pp. 115–120.

225. Fennewald DJ, Weaber RL, Lamberson WR. Genotype by environment interaction for stayability of Red Angus in the United States. Journal of Animal Science. 2018. pp. 422–429. doi:10.1093/jas/skx080

226. Blackburn HD, Krehbiel B, Ericsson SA, Wilson C, Caetano AR, Paiva SR. A fine structure genetic analysis evaluating ecoregional adaptability of a Bos taurus breed (Hereford). PLoS One. 2017;12: e0176474.

227. Hoff JL, Decker JE, Schnabel RD, Seabury CM, Neibergs HL, Taylor JF. QTL-mapping and genomic prediction for bovine respiratory disease in U.S. Holsteins using sequence imputation and feature selection. BMC Genomics. 2019;20: 555.

228. Faux A-M, Gorjanc G, Gaynor RC, Battagin M, Edwards SM, Wilson DL, et al. AlphaSim: Software for Breeding Program Simulation. Plant Genome. 2016;9. doi:10.3835/plantgenome2016.02.0013

229. Chen GK, Marjoram P, Wall JD. Fast and flexible simulation of DNA sequence data. Genome Res. 2009;19: 136–142.

230. Hayes B, Goddard ME. The distribution of the effects of genes affecting quantitative traits in livestock. Genet Sel Evol. 2001;33: 209–229.

231. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. JOSS. 2019;4: 1686.

232. Berg IL, Neumann R, Lam K-WG, Sarbajna S, Odenthal-Hesse L, May CA, et al. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nat Genet. 2010;42: 859–863.

233. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, et al. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science. 2010;327: 836–840.

234. Gonen S, Battagin M, Johnston SE, Gorjanc G, Hickey JM. The potential of shifting recombination hotspots to increase genetic gain in livestock breeding. Genet Sel Evol. 2017;49: 55.

235. Snelling WM, Allan MF, Keele JW, Kuehn LA, McDaneld T, Smith TPL, et al. Genome-wide association study of growth in crossbred beef cattle. J Anim Sci. 2010;88: 837–848.

236. Mateescu RG, Garrick DJ, Reecy JM. Network Analysis Reveals Putative Genes Affecting Meat Quality in Angus Cattle. Front Genet. 2017;8: 171.

237. Mediero A, Cronstein BN. Adenosine and bone metabolism. Trends Endocrinol Metab. 2013;24: 290–300.

238. Alexandre PA, Kogelman LJA, Santana MHA, Passarelli D, Pulz LH, Fantinato-Neto P, et al. Liver transcriptomic networks reveal main biological processes associated with feed efficiency in beef cattle. BMC Genomics. 2015;16: 1073.

239. Li M, Tian S, Jin L, Zhou G, Li Y, Zhang Y, et al. Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. Nat Genet. 2013;45: 1431–1438.

240. Magalhães AFB, de Camargo GMF, Fernandes GA Junior, Gordo DGM, Tonussi RL, Costa RB, et al. Genome-Wide Association Study of Meat Quality Traits in Nellore Cattle. PLoS One. 2016;11: e0157845.

241. Boulant JA, Dean JB. Temperature receptors in the central nervous system. Annu Rev Physiol. 1986;48: 639–654.

242. Mombaerts P. Axonal wiring in the mouse olfactory system. Annu Rev Cell Dev Biol. 2006;22: 713–737.

243. Kellogg DL Jr, Pérgola PE, Piest KL, Kosiba WA, Crandall CG, Grossmann M, et al. Cutaneous active vasodilation in humans is mediated by cholinergic nerve cotransmission. Circ Res. 1995;77: 1222–1228.

244. Kellogg DL Jr, Hodges GJ, Orozco CR, Phillips TM, Zhao JL, Johnson JM. Cholinergic mechanisms of cutaneous active vasodilation during heat stress in cystic fibrosis. J Appl Physiol. 2007;103: 963–968.

245. Joyner MJ, Dietz NM. Sympathetic vasodilation in human muscle. Acta Physiol Scand. 2003;177: 329–336.

246. Smith CJ, Johnson JM. Responses to hyperthermia. Optimizing heat dissipation by convection and evaporation: Neural control of skin blood flow and sweating in humans. Auton Neurosci. 2016;196: 25–36.

247. Takemoto Y. Amino acids that centrally influence blood pressure and regional blood flow in conscious rats. J Amino Acids. 2012;2012: 831759.

248. Meng W, Tobin JR, Busija DW. Glutamate-induced cerebral vasodilation is mediated by nitric oxide through N-methyl-D-aspartate receptors. Stroke. 1995;26: 857–62; discussion 863.

249. Conrad KP. Unveiling the vasodilatory actions and mechanisms of relaxin. Hypertension. 2010;56: 2–9.

250. Daanen HAM, Van Marken Lichtenbelt WD. Human whole body cold adaptation. Temperature (Austin). 2016;3: 104–118.

251. Choshniak I, McEwan-Jenkinson D, Blatchford DR, Peaker M. Blood flow and catecholamine concentration in bovine and caprine skin during thermal sweating.

Comp Biochem Physiol C. 1982;71C: 37–42.

252. Rhoads ML, Rhoads RP, VanBaale MJ, Collier RJ, Sanders SR, Weber WJ, et al. Effects of heat stress and plane of nutrition on lactating Holstein cows: I. Production, metabolism, and aspects of circulating somatotropin. J Dairy Sci. 2009;92: 1986–1997.

253. El-Nouty FD, Elbanna IM, Davis TP, Johnson HD. Aldosterone and ADH response to heat and dehydration in cattle. J Appl Physiol. 1980;48: 249–255.

254. Itoh F, Obara Y, Rose MT, Fuse H, Hashimoto H. Insulin and Glucagon Secretion in Lactating Cows During Heat Exposure1. Available: https://academic.oup.com/jas/article-abstract/76/8/2182/4643238

255. Sanz Fernandez MV, Stoakes SK, Abuajamieh M, Seibert JT, Johnson JS, Horst EA, et al. Heat stress increases insulin sensitivity in pigs. Physiol Rep. 2015;3. doi:10.14814/phy2.12478

256. Keogh K, Kenny DA, Kelly AK, Waters SM. Insulin secretion and signaling in response to dietary restriction and subsequent re-alimentation in cattle. Physiol Genomics. 2015;47: 344–354.

257. Tesmer LA, Lundy SK, Sarkar S, Fox DA. Th17 cells in human disease. Immunol Rev. 2008;223: 87–113.

258. Romagnani S. T-cell subsets (Th1 versus Th2). Ann Allergy Asthma Immunol. 2000;85: 9–18; quiz 18, 21.

259. Deschamps M, Laval G, Fagny M, Itan Y, Abel L, Casanova J-L, et al. Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. Am J Hum Genet. 2016;98: 5–21.

260. Crawford JE, Amaru R, Song J, Julian CG, Racimo F, Cheng JY, et al. Natural Selection on Genes Related to Cardiovascular Health in High-Altitude Adapted Andeans. Am J Hum Genet. 2017;101: 752–767.

261. Sunkin SM, Ng L, Lau C, Dolbeare T, Gilbert TL, Thompson CL, et al. Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. Nucleic Acids Res. 2013;41: D996–D1008.

262. Kang EY, Han B, Furlotte N, Joo JWJ, Shih D, Davis RC, et al. Meta-analysis identifies gene-by-environment interactions as demonstrated in a study of 4,965 mice. PLoS Genet. 2014;10: e1004022.

263. Bradley DG, MacHugh DE, Cunningham P, Loftus RT. Mitochondrial diversity and the origins of African and European cattle. Proc Natl Acad Sci U S A. 1996;93: 5131–5135.

264. Kemper KE, Saxton SJ, Bolormaa S, Hayes BJ, Goddard ME. Selection for complex traits leaves little or no classic signatures of selection. BMC Genomics. 2014;15: 246.

265. Zhang F, Wang Y, Mukiibi R, Chen L, Vinsky M, Plastow G, et al. Genetic architecture of quantitative traits in beef cattle revealed by genome wide association studies of imputed whole genome sequence variants: I: feed efficiency and component traits. BMC Genomics. 2020;21: 36.

266. Mackay TFC, Stone EA, Ayroles JF. The genetics of quantitative traits: challenges and prospects. Nat Rev Genet. 2009;10: 565–577.

267. Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. Genetic architecture: the shape of the genetic contribution to human traits and disease. Nat Rev Genet. 2018;19: 110–124.

268. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42: 565–569.

269. Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012;44: 369–75, S1–3.

270. Fang L, Cai W, Liu S, Canela-Xandri O, Gao Y, Jiang J, et al. Comprehensive analyses of 723 transcriptomes enhance genetic and biological interpretations for complex traits in cattle. Genome Res. 2020;30: 790–801.

271. Durkin K, Coppieters W, Drögemüller C, Ahariz N, Cambisano N, Druet T, et al. Serial translocation by means of circular intermediates underlies colour sidedness in cattle. Nature. 2012;482: 81–84.

272. He M, Cornelis MC, Kraft P, van Dam RM, Sun Q, Laurie CC, et al. Genome-wide association study identifies variants at the IL18-BCO2 locus associated with interleukin-18 levels. Arterioscler Thromb Vasc Biol. 2010;30: 885–890.

273. Gao H, Wu Y, Li J, Li H, Li J, Yang R. Forward LASSO analysis for high-order interactions in genome-wide association study. Brief Bioinform. 2014;15: 552–561.

274. Braz CU, Taylor JF, Bresolin T, Espigolan R, Feitosa FLB, Carvalheiro R, et al. Sliding window haplotype approaches overcome single SNP analysis limitations in identifying genes for meat tenderness in Nelore cattle. BMC Genet. 2019;20: 8.

275. Vereecke L, Beyaert R, van Loo G. The ubiquitin-editing enzyme A20 (TNFAIP3) is a central regulator of immunopathology. Trends Immunol. 2009;30: 383–391.

276. Graham RR, Cotsapas C, Davies L, Hackett R, Lessard CJ, Leon JM, et al. Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus. Nat Genet. 2008;40: 1059–1061.

277. Doherty R, Whiston R, Cormican P, Finlay EK, Couldrey C, Brady C, et al. The CD4(+) T cell methylome contributes to a distinct CD4(+) T cell transcriptional signature in Mycobacterium bovis-infected cattle. Sci Rep. 2016;6: 31014.

278. Jensen K, Paxton E, Waddington D, Talbot R, Darghouth MA, Glass EJ. Differences in the transcriptional responses induced by Theileria annulata infection in bovine monocytes derived from resistant and susceptible cattle breeds. Int J Parasitol. 2008;38: 313–325.

279. Khatib H, Zaitoun I, Kim E-S. Comparative analysis of sequence characteristics of imprinted genes in human, mouse, and cattle. Mamm Genome. 2007;18: 538–547.

280. Giuffra E, Tuggle CK, FAANG Consortium. Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap. Annu Rev Anim Biosci. 2019;7: 65–88.

281. Purfield DC, Evans RD, Berry DP. Breed- and trait-specific associations define the genetic architecture of calving performance traits in cattle. J Anim Sci. 2020;98. doi:10.1093/jas/skaa151

282. Xiang R, van den Berg I, MacLeod IM, Hayes BJ, Prowse-Wilkins CP, Wang M, et al. Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. Proc Natl Acad Sci U S A. 2019;116: 19398–19408.

283. Islam R, Liu X, Gebreselassie G, Abied A, Ma Q, Ma Y. Genome-wide association analysis reveals the genetic locus for high reproduction trait in Chinese Arbas Cashmere goat. Genes Genomics. 2020;42: 893–899.

284. Cole JB, Wiggans GR, Ma L, Sonstegard TS, Lawlor TJ Jr, Crooker BA, et al. Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. BMC Genomics. 2011;12: 408.

285. Paim T do P, Hay EHA, Wilson C, Thomas MG, Kuehn LA, Paiva SR, et al. Genomic Breed Composition of Selection Signatures in Brangus Beef Cattle. Front Genet. 2020;11: 710.

286. Pérez O’Brien AM, Utsunomiya YT, Mészáros G, Bickhart DM, Liu GE, Van Tassell CP, et al. Assessing signatures of selection through variation in linkage disequilibrium between taurine and indicine cattle. Genet Sel Evol. 2014;46: 19.

287. Sartorelli V, Lauberth SM. Enhancer RNAs are an important regulatory layer of

the epigenome. Nat Struct Mol Biol. 2020;27: 521–528.

288. Orozco-terWengel P, Kapun M, Nolte V, Kofler R, Flatt T, Schlötterer C. Adaptation of Drosophila to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. Mol Ecol. 2012;21: 4931–4941.

289. Beissinger TM, Rosa GJM, Kaeppler SM, Gianola D, de Leon N. Defining window-boundaries for genomic analyses using smoothing spline techniques. Genet Sel Evol. 2015;47: 30.

290. Hayes BJ, Daetwyler HD. 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. Annu Rev Anim Biosci. 2018. doi:10.1146/annurev-animal-020518-115024

291. Lindholm-Perry AK, Kuehn LA, Smith TPL, Ferrell CL, Jenkins TG, Freetly HC, et al. A region on BTA14 that includes the positional candidate genes LYPLA1, XKR4 and TMEM68 is associated with feed intake and growth phenotypes in cattle(1). Anim Genet. 2012;43: 216–219.

292. Fortes MRS, Reverter A, Kelly M, McCulloch R, Lehnert SA. Genome-wide association study for inhibin, luteinizing hormone, insulin-like growth factor 1, testicular size and semen traits in bovine species. Andrology. 2013;1: 644–650.

293. Zhang R, Miao J, Song Y, Zhang W, Xu L, Chen Y, et al. Genome-wide association study identifies the PLAG1-OXR1 region on BTA14 for carcass meat yield in cattle. Physiol Genomics. 2019;51: 137–144.

294. Chen M-H, Raffield LM, Mousas A, Sakaue S, Huffman JE, Moscati A, et al. Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. Cell. 2020;182: 1198–1213.e14.

295. Zhang B, Peñagaricano F, Driver A, Chen H, Khatib H. Differential expression of heat shock protein genes and their splice variants in bovine preimplantation embryos. J Dairy Sci. 2011;94: 4174–4182.

296. Cochran SD, Cole JB, Null DJ, Hansen PJ. Single nucleotide polymorphisms in candidate genes associated with fertilizing ability of sperm and subsequent embryonic development in cattle. Biol Reprod. 2013;89: 69.

297. Comuzzie AG, Cole SA, Laston SL, Voruganti VS, Haack K, Gibbs RA, et al. Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. PLoS One. 2012;7: e51954.

298. Gutiérrez-Gil B, Wiener P, Williams JL. Genetic effects on coat colour in cattle: dilution of eumelanin and phaeomelanin pigments in an F2-Backcross Charolais x Holstein population. BMC Genet. 2007;8: 56.

299. Rosenberg NA, Edge MD, Pritchard JK, Feldman MW. Interpreting polygenic

scores, polygenic adaptation, and human phenotypic differences. Evol Med Public Health. 2019;2019: 26–34.

300. Nei M, Tajima F. Genetic drift and estimation of effective population size. Genetics. 1981;98: 625–640.

301. Chan EKF, Rowe HC, Corwin JA, Joseph B, Kliebenstein DJ. Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in Arabidopsis thaliana. PLoS Biol. 2011;9: e1001125.

302. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017;169: 1177–1186.

303. Liu X, Li YI, Pritchard JK. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell. 2019;177: 1022–1034.e6.

304. Beissinger T, Kruppa J, Cavero D, Ha N-T, Erbe M, Simianer H. A Simple Test Identifies Selection on Complex Traits. Genetics. 2018;209: 321–333.

305. Turchin MC, Chiang CWK, Palmer CD, Sankararaman S, Reich D, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, et al. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nat Genet. 2012;44: 1015–1019.

306. Szpiech ZA, Novak TE, Bailey NP, Stevison LS. High-altitude adaptation in rhesus macaques. 2020. p. 2020.05.19.104380. doi:10.1101/2020.05.19.104380

307. Hu Z-L, Park CA, Reecy JM. Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB. Nucleic Acids Res. 2019;47: D701–D710.

308. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20: 1297–1303.

309. Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol Biol Evol. 2014;31: 1275–1291.

310. Szpiech ZA, Hernandez RD. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. Mol Biol Evol. 2014;31: 2824–2827.

311. Alachiotis N, Pavlidis P. RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. Commun Biol. 2018;1: 79.

312. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al.

Ensembl 2020. Nucleic Acids Res. 2020;48: D682–D688.

313. Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Brief Bioinform. 2013;14: 144–161.

VITA

Troy Neal Rowan was born September 3, 1993 in Bedford, Iowa. From a young age, Troy was immediately taken with his family's purebred Charolais cows. He spent much of his youth tagging along to custom artificial insemination projects with his father, Kurt. By 4th grade, Troy was an active member in the Taylor County Iowa 4-H program, first as a member of the Lucky 4 club, then as a founding member of the Ross Wranglers. He successfully exhibited home-raised animals at the county, state, and national level. Throughout elementary, middle, and high school Troy attended the Charolais Junior National Show and Conference. There he exhibited home-raised heifers and steers and participated in a variety of public speaking contests.

After graduating from Bedford Community High School in 2012, Troy attended Creighton University in Omaha, Nebraska where he majored in Biology. His initial intent was to follow in his mother Theresa's footsteps and become a pharmacist. Early on though, he began undergraduate research in Dr. Karin van Dijk's lab, first studying epigenetic modifications due to bacterial infections in *Arabidopsis thaliana*, then studying algal lipid production under the supervision of Dr. Mike McConnell. After Dr. van Dijk's lab moved to the University of Nebraska, Troy began work with Dr. Carol Fassbinder-Orth, studying viral dynamics in an arthropod-avian system.

These research experiences, and an interest in returning to the agricultural community led Troy to a Ph.D. at the University of Missouri as a part of their Genetics Area Program. Troy began his Ph.D. work with Dr. Jared Decker in the Fall of 2016. During his time at Mizzou, Troy used commercially-generated genotypes from beef cattle to understand the population genetics phenomena of polygenic selection and local adaptation. Troy published multiple peer-reviewed journal articles and delivered podium presentations at four international genetics conferences. Additionally, Troy took an active role in Extension programming, delivering 12 presentations to cattle producers and allied industry professionals.